Computer intensive methods for evaluating latent class model fit

van Kollenburg, Geert

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Computer Intensive Methods for Evaluating Latent Class Model Fit

Proefschrift

ter verkrijging van de graad van doctor
aan Tilburg University
op gezag van de rector magnificus,
prof. dr. E.H.L. Aarts,
in het openbaar te verdedigen ten overstaan van een
door het college voor promoties ingestelde commissie
in de aula van de Universiteit

op maandag 13 november 2017 om 16.00 uur

door

Geert Hein van Kollenburg,

Promotor:	Prof. dr. J.K. Vermunt
Copromotor:	Dr. ir. J. Mulder
Overige Leden:	Dr. S. Bouwmeester
	Prof. dr. H. Hoijtink
	Dr. D.L. Oberski
	Prof. dr. A.G. de Waal

Contents

Chapter 1

Introduction

1.1 What is Latent Class Analysis?

The basis of this dissertation is the latent class (LC) model (Goodman, 1974). This is a powerful tool with which observations can be clustered based on their combination of responses to a number of categorical variables. Categorical variables are those which are scored as yes/no, agree/neutral/disagree, or symptom present/absent and so on. The LC model is widely used in many research fields such as psychiatry (Roedelof, Bongers, & van Nieuwenhuizen, 2013), abnormal psychology (Crow et al., 2012), biomedical sciences (Rindskopf, 2002), developmental psychology (Laudy et al., 2005), marketing (Okazaki, Campo, Andreu, & Romero, 2014) and gambling studies (Dufour, Brunelle, & Roy, 2013).

To conceptually introduce the LC model itself, let us start with the basic independence model for categorical data analysis (Agresti, 2002). Say that we have N observations on a number of categorical variables. Those variables combined can make different response patterns. For example, when we have J dichotomous variables (with two possible categories), there are $S = 2^J$ possible response patterns comprising all possible combinations of

J ones and zeros. When the variables are independent (as the independence model assumes), the probability of observing a particular pattern is equal to the product of the individual response probabilities.

But what if the N observations actually belong to different groups (or, classes) which differ with respect to response probabilities to the variables? For example, we may expect men and women to have differing beliefs on some variables. Extending the independence model to accommodate multiple groups changes the main assumption of independence to local (or, conditional) independence within each group. That is, rather than overall independence, we now assume that within each group (i.e., the men and women) the variables are independent. If we do not take into account that the observations belong to different groups we may still observe associations between the variables.

Now, the LC model formulation is the same as the conditional independence model just described. There is, however, a rather crucial difference. We use the LC model when we do not know which observation belongs to which group, or how many groups there actually are. That is, the variable which would normally indicate to which group an observation belongs is now a latent (unobserved) variable. We use the LC model to try to cluster the respondents based on their response patterns. The resulting clustering may not be as clear-cut as the male/female distinction, however.

1.2 Purpose of this Research

An important part of using statistical models is assessing whether that model is in agreement with the observed data. This is usually done by calculating a statistic which compares the observed data with model predictions. Whether the value of a statistic indicates model misfit is determined

by means of a p value. A p value gives the probability that we would find the observed value for the statistic, or more extreme, if the model under consideration generated the observed data. A very low probability (close to 0) means that it is very unlikely that the data came from the model. This implies that the model may be incorrect.

The research leading to this dissertation focussed on different types of p values for various statistics in the context of LC analysis. Besides only the comparison of different methods, the conducted research led to surprising integrations of frequentist and Bayesian methods. By doing so, some of the issues related to existing model fit testing methods were solved. The topics in the conducted research were:

1. Compare different p values for the most commonly used statistics in LC analysis.

2. Modify the Bayesian posterior predictive p value, by improving its frequentist properties.

3. Reduce the computational burden of frequentist resampling methods, by incorporating Bayesian ideas.

1.3 Outline of the Dissertation

In the next chapter, the frequentist properties (type I error rates and power) of different p values were evaluated for many commonly used statistics in LC analysis. In this chapter we discuss algorithms to calculate asymptotic, parametric bootstrap and two kinds of posterior predictive p values for a number of commonly used statistics for LC model fit testing. Various Monte Carlo simulation studies provided a direct comparison of type I error rates and power of the different p values. The results led to a

number of recommendations and warnings for certain (combinations of) p values and statistics.

In the subsequent chapter a calibration method of the posterior predictive p value is discussed. The resulting calibrated p value is easier to interpret and results in higher power to reject a misspecified model than the original posterior predictive p value. A generic algorithm is developed and its application was illustrated in the context of LC models and linear regression analysis. The Monte Carlo studies showed that indeed the type I error rates and power of the new calibrated p value were superior to the original posterior predictive p value.

The final chapter of this dissertation provides a solution to the computational burden of current frequentist resampling methods to obtain p values for LC models. Where current methods require a model of interest to be estimated hundreds of times, the newly proposed methodology requires a model to be estimated only once. This speeds up the calculation of a p value tremendously. For it to work, it borrows an idea from the Bayesian paradigm. That is, it compares observed data directly with many model-generated data sets on aspects we as researchers find important. If the model produces data which is very different from what we have observed in the data, we can certain that the model did not produce the observed data. A Monte Carlo study showed that the method had very low probability to falsely reject a model that produced the observed data (i.e., it had low type I error rates). Illustration with an empirical data set showed that the new methodology resulted in the same conclusions as the computationally much more demanding parametric bootstrap procedure.

The chapters were written as separate articles to be published in academic journals. They are kept as close as possible to the original manuscripts and contain some overlap content-wise. Some notation also differs.

Chapter 2

Assessing Model fit in Latent Class analysis when Asymptotics do not hold

Abstract

The application of latent class (LC) analysis involves evaluating the LC model using goodness-of-fit statistics. To assess the misfit of a specified model, say with the Pearson chi-squared statistic, a p value can be obtained using an asymptotic reference distribution. However, asymptotic p values are not valid when the sample size is not large and/or the analysed contingency table is sparse. Another problem is that for various other conceivable global and local fit measures, asymptotic distributions are not readily available. An alternative way to obtain the p value for the statistic of interest is by constructing its empirical reference distribution using

This chapter is published as van Kollenburg. G.H., Mulder, J., & Vermunt, J.K. (2015). Assessing Model fit in Latent Class analysis when Asymptotics do not hold. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *11*(2), 65–79.

resampling techniques such as the parametric bootstrap or the posterior predictive check (PPC). In the current chapter, we show how to apply the parametric bootstrap and two versions of the PPC to obtain empirical p values for a number of commonly used global and local fit statistics within the context of LC analysis. The main difference between the model-based PPC and the parametric bootstrap is that the former takes into account parameter uncertainty. The parameter-based PPC has the advantage that it is computationally much less intensive than the other two resampling methods.

In a Monte Carlo study we evaluated type I error rates and power of these resampling methods when used for global and local goodness-of-fit testing in LC analysis. Results show that both the bootstrap and the model-based PPC are generally good alternatives to asymptotic p values and can also be used when (asymptotic) distributions are not known. Nominal type I error rates were not met when sample size was small and the contingency table has many cells. Overall the model-based PPC was somewhat more conservative than the parametric bootstrap. We have also replicated previous research suggesting that the Pearson X^2 statistic should in many cases be preferred over the likelihood-ratio G^2 statistic. Power to reject a model for which the number of LCs was 1 lower than in the population was very high, unless sample size was small. When the contingency tables are very sparse, the $TBVR$ statistic, which is based on bivariate relationships, still had very high power, signifying its usefulness in assessing model fit.

2.1 Introduction

The use of latent class (LC) models is becoming more and more widespread in a broad range of fields, such as in biomedical sciences (Rindskopf, 2002), psychiatry (Roedelof et al., 2013), abnormal psychology (Crow et al., 2012) developmental psychology (Laudy et al., 2005), gambling studies (Dufour et al., 2013) and marketing (Okazaki et al., 2014). This makes the availability of reliable methods to assess the goodness-of-fit of LC models in-

creasingly important (Lanza, Flaherty, & Collins, 2004).

The global or overall goodness-of-fit of a LC model is typically assessed using the Pearson or the likelihood-ratio chi-squared statistic (Goodman, 1974). For local fit assessment, which involves checking whether the specified LC model describes specific aspects of the data well, various types of statistics have been proposed, such as residual log-odds-ratios and Pearson statistics computed in two-way tables (Hagenaars, 1988; Magidson & Vermunt, 2004). A convenient way to determine the extent of global or local misfit is to obtain p values for the goodness-of-fit statistics of interest. Typically, we would get the p values from the asymptotic distributions of the statistics, but these are not always readily available. Moreover, even when these are available, asymptotic p values are not useful when the analysed contingency table is too sparse because the sample size is small or the number of cells in the table is large (Haberman, 1988; Langeheine, Pannekoek, & Van de Pol, 1996; Maydeu-Olivares & Joe, 2006; Reiser & Lin, 1999).

We can also obtain p values by using resampling techniques, such as the parametric bootstrap (Efron & Tibshirani, 1993) or the posterior predictive check (PPC) (Meng, 1994; Rubin, 1984). The major benefit of resampling techniques over asymptotics is that we do not need any distributional assumptions regarding the statistics. These methods generate replicated data sets based on the parameter estimates for the specified model, and for each data set they calculate the required statistics. The p values for the statistics are then obtained from their resulting empirical distributions. The main difference between the model-based PPC and the parametric bootstrap is that the former takes into account parameter uncertainty. Another variant of the PPC called the parameter-based PPC has the advantage that it is computationally much less intensive than the

other two resampling methods.

While bootstrap methods have been proposed in the context of LC analysis as a way to deal with sparseness when assessing global fit (Langeheine et al., 1996; Von Davier, 1997), they have not been used so far to obtain p values for statistics for which the distributions are unknown, such as the local fit measures proposed by Magidson and Vermunt (2004). In contrast, PPCs have been used to assess LC model fit using a range of global and local fit measures (Berkhof, Van Mechelen, & Gelman, 2003; Hoijtink, 1998; Ligtvoet & Vermunt, 2012; Meulders, De Boeck, Kuppens, & Van Mechelen, 2002; Rubin & Stern, 1994), but the performance of this approach has not been investigated in a systematic manner.

The purpose of this chapter is to discuss and investigate bootstrap and PPC methods in a more integrated manner. This allows expanding the bootstrapping approach to obtain p values not only in case of sparseness, but also with measures for which the asymptotic distribution is unknown. This allows answering the question as to whether the PPC can be an improvement over the bootstrap when the latter works less well (Von Davier, 1997). More specifically, does taking parameter uncertainty into account yield more reliable p values when tables are extremely sparse?

The remainder of this chapter is organised as follows. Section 2.2 reviews the LC model and describes a number of commonly used statistics to asses global and local LC model fit. In Section 2.3 we discuss the various methods to obtain p values in more detail. Section 2.4 presents a simulation experiment in which the performance of the investigated methods to obtain p values is compared. In Section 2.5 we present an empirical example and finally in Section 2.6 we discuss the main findings and issues in need of further research.

2.2 Latent Class Analysis

2.2.1 The Model

Suppose we have N observations on J categorical variables with R_j categories for variable number j ($j = 1, \ldots, J$). There are then $S = \prod_{j=1}^{J} R_j$ possible response patterns, which can be denoted as $\boldsymbol{y}_s = (y_{s1}, \ldots, y_{sJ}), s = 1, \ldots, S$. Letting n_s denote the observed frequency for pattern $\boldsymbol{y}_s$, the observed data can be summarised as pattern frequencies in $\mathbf{n} = (n_1, \ldots, n_S)$.

The LC model assumes that the N observations can be partitioned into C latent classes, which form the categories of the discrete latent variable ξ (Goodman, 1974). The LCs differ from one another with respect to the conditional response probabilities to the variables. Moreover, within each LC the responses to the observed variables are assumed to be independent of one another (i.e., the local independence assumption).

Let ρ_c be the class size (proportion) of LC c and let π_{rjc} be the conditional response probability that a respondent gives response r to variable j, given that he or she belongs to LC c. The probability of observing response pattern s is then a mixture of multinomial distributions with weights equal to the class proportion ρ_c. It is given by:

$$P(\boldsymbol{y}_s) = \sum_{c=1}^{C} \rho_c \prod_{j=1}^{J} \prod_{r=1}^{R_j} \pi_{rjc}^{y^*_{sjr}}, \tag{2.1}$$

where y^*_{sjr} is 1 if $y_{sj} = r$ and 0 otherwise.

Several methods exist to estimate the LC model parameters $\boldsymbol{\theta} = (\boldsymbol{\rho}, \boldsymbol{\pi})$. One might be interested in obtaining point estimates, interval estimates or posterior probability distributions for the unknown parameters. To obtain their maximum likelihood estimates we typically use the expectation-

maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). We may obtain estimates of their posterior distribution by means of an MCMC algorithm (Tanner & Wong, 1987).

2.2.2 Goodness-of-Fit Measures

An important part of the model selection procedure in LC modeling involves checking whether a model is in agreement with the data. The discrepancies between observed data and expectations under the model can be assessed using goodness-of-fit (GoF) statistics (Agresti, 2002). We will discuss statistics for the assessment of global and local fit.

Global fit statistics aggregate the disagreement between the observed frequencies n_s and the expected frequencies under the model $e_s = N \cdot P(\mathbf{y}_s)$ into a single value. Well-known chi-squared statistics are the Pearson X^2,

$$X^2(\mathbf{n}) = \sum_{s=1}^{S} \frac{(n_s - e_s)^2}{e_s}, \tag{2.2}$$

and the likelihood ratio statistic G^2,

$$G^2(\mathbf{n}) = 2 \sum_{s=1}^{S} n_s \ln(n_s/e_s). \tag{2.3}$$

These two chi-squared statistics belong to the more general family of power divergence statistics which take the form

$$PD(\mathbf{n}) = \frac{2}{\lambda(\lambda+1)} \sum_{s=1}^{S} n_s \left\{ \left(\frac{n_s}{e_s} \right)^{\lambda} - 1 \right\}. \tag{2.4}$$

The X^2 and G^2 statistics are obtained by setting $\lambda = 1$ and letting λ approach 0, respectively. These two statistics have been shown to be in-

appropriate when contingency tables are sparse; that is, when a portion of the expected frequencies is small. In such cases, the X^2 statistic tends to become very large, yielding a p value of 0, while the G^2 statistic tends to be small, yielding a p value of 1. It has been argued that a good trade-off is found by setting λ equal to 2/3, through which we obtain the Cressie-Read (CR) statistic (Cressie & Read, 1984).

Another global fit measure indicative of how much the observed and estimated cell frequencies differ is the Dissimilarity Index (DI):

$$DI(\mathbf{n}) = \frac{\sum_{s=1}^{S} |n_s - e_s|}{2N}. \tag{2.5}$$

The DI indicates which proportion of the sample should be moved to another cell to obtain a perfect fit (Vermunt & Magidson, 2013). Though this statistic is appealing due to the information it provides, its asymptotic distribution is unknown. Therefore, to obtain a p value for this statistics, we need to resort to resampling techniques.

In LC modeling, local fit is typically assessed by computing statistics for lower-order marginals of the analysed J-way contingency table. A popular and very useful measure is the bivariate residual (BVR) statistic, which can be used to determine violations of the local independence assumption (Magidson & Vermunt, 2004; Vermunt & Magidson, 2013). The BVR quantifies the residual association between pairs of variables using a Pearson-like chi-squared statistics. To show how the BVR is calculated, let the subscript r indicate a given response to variable j and subscript $r^{'}$ a response to variable $j^{'}$. Then $n_{rr'}$ indicates an observed frequency in the two-way cross-tabulation of variables j and $j^{'}$. The expected frequency for this pattern, $e_{rr'}$, can be calculated from the LC model parameters as

follows:

$$e_{rr'} = N \sum_{c=1}^{C} \rho_c \pi_{rjc} \pi_{r'j'c}.$$

The BVR for the variable pair j-$j^{'}$ is then:

$$BVR_{jj'}(\mathbf{n}) = \sum_{r=1}^{R_j} \sum_{r'=1}^{R_{j'}} \frac{(n_{rr'} - e_{rr'})^2}{e_{rr'}}. \tag{2.6}$$

Similar Pearson-like local fit measures may be computed for higher-order tables, for example, for cross-tabulations of three instead of two variables. An important advantage of the BVR statistic compared to global fit measures is that it is much less sensitive to sparseness (Maydeu-Olivares & Joe, 2006). A disadvantage is, however, that its asymptotic distribution is not known, implying that asymptotic p values are not available.

Based on the BVR, we can derive a global fit measure that may be used as an alternative to the standard GoF chi-squared statistics. This total BVR ($TBVR$) statistic is obtained by summing the BVR statistics across all variable pairs, that is,

$$TBVR(\mathbf{n}) = \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} BVR_{jj'}(\mathbf{n}). \tag{2.7}$$

The main advantage of the $TBVR$ is that is much less affected by sparseness than other global fit measures. However, as for the BVRs themselves, also for the $TBVR$ the asymptotic distribution is unknown. And although knowledge on lower-order fit is very useful, we cannot rule out higher-order misfit, due to multivariate interactions, based on lower-order statistics (Reiser & Lin, 1999).

2.3 Determining p Values for GoF Measures

2.3.1 Asymptotic p Values

To test whether a model deviates from the data, most often a p value is calculated based on an asymptotic reference distribution. If a C-class model is true, the power-divergence statistics asymptotically (as N goes to infinity) follow a chi-squared (χ^2_{df}) distribution with degrees of freedom (df) equal to

$$df = \prod_{j=1}^{J} R_j - C(1 + \sum_{j=1}^{J}(R_j - 1)) \tag{2.8}$$

(Haberman, 1979; Magidson & Vermunt, 2004). The p value is then equal to the tail-area probability that a value from the χ^2_{df} distribution is equal to or greater than the computed statistic. If the p value is less than some a priori set threshold (e.g., .05), the researcher concludes that there is significant misfit between the model and the data (Fisher, 1925).

An important issue related to the use of asymptotic reference distributions is that it is not accurate when the corresponding frequency table is sparse. This occurs when the sample size is not large enough for the contingency table at hand. For example, 10 dichotomous variables create a table with $2^{10} = 1024$ cells, which would be considered sparse even with 1000 observations. Sparse tables result in untrustworthy asymptotic p values (see e.g., Collins, Fidler, Wugalter, & Long, 1993; Langeheine et al., 1996; Magidson & Vermunt, 2004; Von Davier, 1997).

For statistics such as the DI, BVR, and $TBVR$, asymptotic distributions are not known. In some cases rules of thumb are used, but these may not always be accurate. For instance, one rule of thumb says that for dichotomous variables BVR values greater than 3.84 indicate significant misfit (3.84 being the 95th percentile of the χ^2_1 distribution). Others take

BVR values greater than 1 to indicate misfit. It appears that each cut-off has its down-sides and can result in too conservative or too liberal conclusions, depending on the situation (Oberski, van Kollenburg, & Vermunt, 2013). Resampling techniques are therefore required to obtain p values.

2.3.2 Parametric Bootstrap

To overcome the problems associated with asymptotic p values it is possible to obtain empirical reference distributions through resampling methods like the parametric bootstrap (Efron & Tibshirani, 1993), which is used in LC analysis regularly (see e.g., Formann, 2003; Jansen & van der Maas, 1997; Lin, McCulloch, Turnbull, Slate, & Clark, 2000). The parametric bootstrap simulates the probability of finding a value for a statistic T, greater than or equal to the observed value of the statistic $T(\mathbf{n})$, conditional on the ML estimates for the C-class model being the population parameters. The parametric bootstrap p value for a statistic T is obtained as follows:

Algorithm 2.1: Parametric Bootstrap in LC Analysis

Step 1: Find the ML estimates $\hat{\boldsymbol{\theta}}$ for the C-class model (for instance using EM) and calculate the observed fit-statistic $T(\mathbf{n}^{\text{obs}})$. For example, one could use the Pearson X^2 statistic, in which case $T(\mathbf{n}^{\text{obs}}) = X^2(\mathbf{n}^{\text{obs}})$.

Step 2: Calculate the estimated pattern probabilities $\widehat{P}(\mathbf{y}_s)$ from the ML estimates $\hat{\boldsymbol{\theta}}$. Draw B random replicated samples, $\mathbf{n}^{\text{rep}}$, of size N from a multinomial distribution with parameters $\widehat{P}(\mathbf{y}_s)$:

$$\mathbf{n}^{\text{rep},(b)} \sim \text{Multin}(N, \widehat{P}(\mathbf{y}_1), \ldots, \widehat{P}(\mathbf{y}_S)), b = 1, \ldots, B \tag{2.9}$$

Step 3: Determine the empirical reference distribution of the statistic $T(\mathbf{n})$. That is, find the ML estimates for each data set $\mathbf{n}^{\text{rep},(b)}$ and calculate $T(\mathbf{n}^{\text{rep},(b)})$. For instance, calculate $X^2(\mathbf{n}^{\text{rep},(b)})$ (and/or other statistics of interest).

Step 4: Estimate the bootstrap p value by the proportion of $T(\mathbf{n}^{\text{rep},(b)})$ which are greater than, or equal to $T(\mathbf{n})$ (which was calculated in Step 1):

$$\hat{p}_{\text{boot}} = B^{-1} \sum_{b=1}^{B} I(T(\mathbf{n}^{\text{rep},(b)}) \geqslant T(\mathbf{n}^{\text{obs}})), \tag{2.10}$$

where the indicator function I equals 1 if $T(\mathbf{n}^{\text{rep},(b)}) \geqslant T(\mathbf{n}^{\text{obs}})$ and 0 otherwise. If $\hat{p}_{\text{boot}}$ is less than a predefined value (for instance, .05) we conclude that the model does not fit the data properly.

Langeheine et al. (1996) showed that the parametric bootstrap method works well with global chi-squared statistics for small well-filled contingency tables. However, Von Davier (1997) showed that in sparse contingency tables with many cells, different conclusions about LC model fit might be obtained depending on which statistic is used. Bootstrap p values for the G^2 statistics were shown to lead to conservative results, while p values for the Pearson's X^2 and CR statistics did not fail systematically. Thus, although we can obtain empirical distributions for any statistic, sparseness can still have an effect on how reliable the resulting p values are, depending on the statistic that is used.

The bootstrap has not been used so far to obtain p values for GoF measures for which asymptotic p values are not available, such as the DI, BVR, and $TBVR$ statistics. Whether the bootstrap is suitable for use with these measures has yet to be determined.

2.3.3 Model-Based PPC

In the parametric bootstrap, each of the B replicated data sets is generated using the same ML estimates as if it were population parameter values, implying that the uncertainty about these estimates is not taken into account. Within the Bayesian framework, parameter uncertainty is incorporated in the posterior distribution. The PPC can be seen as the Bayesian counterpart of the parametric bootstrap which makes use of this posterior distribution.

Two versions of the PPC exist. A model-based PPC and a parameter-based PPC. The model-based PPC which was used in LC analysis by Rubin and Stern (1994), is very similar to the parametric bootstrap. It generates a large number of replicated data sets, re-estimates the LC model for each data set, and calculates the statistics of interest. The only difference is that the model-based PPC uses parameter draws from their posterior distributions as population values to sample the replicated data sets, rather than fixing the parameters to their ML estimates. The parameter-based PPC (Gelman, Meng, & Stern, 1996) does not require re-estimating the LC model for each replicated data set. Instead, it compares both the observed and replicated data directly to the parameter values sampled from their posterior distribution. The parameter-based PPC will be discussed in detail in the next subsection.

In LC analysis, the model-based PPC to obtain a p value for a statistic $T(\mathbf{n})$ (which is based on ML estimates) proceeds as follows:

Algorithm 2.2: Model-based PPC in LC Analysis

Step 1: Find the ML estimates $\hat{\boldsymbol{\theta}}$ for the C-class model (for instance using EM) and calculate the observed fit-statistic $T(\mathbf{n}^{\text{obs}})$. For example, one could use the Pearson X^2 statistic, in which case $T(\mathbf{n}^{\text{obs}}) =$

$X^2(\mathbf{n}^{\text{obs}})$.

Step 2: Obtain K draws $\boldsymbol{\theta}^{(k)}$ from the posterior distribution for the C-class model:

$$\boldsymbol{\theta}^{(k)} \sim p(\boldsymbol{\theta}|\mathbf{n}^{\text{obs}}), k = 1, \ldots, K. \tag{2.11}$$

This can be done using an MCMC algorithm (Rubin & Stern, 1994).

Step 3: Calculate the estimated pattern probabilities $\widehat{P}(\mathbf{y}_s)^{(k)}$ from $\boldsymbol{\theta}^{(k)}$. Draw K random samples of size N from a multinomial distribution with parameters $\widehat{P}(\mathbf{y}_s)^{(k)}$

$$\mathbf{n}^{\text{rep},(k)} \sim \text{Multin}(N, \widehat{P}(\mathbf{y}_1)^{(k)}, \ldots, \widehat{P}(\mathbf{y}_S)^{(k)}) \tag{2.12}$$

Step 4: Obtain the ML estimates (e.g., using the EM algorithm) for each data set $\mathbf{n}^{\text{rep},(k)}$ and calculate $T(\mathbf{n}^{\text{rep},(k)})$ to determine the empirical reference distribution of the statistic T. For instance, calculate $T(\mathbf{n}^{\text{rep},(k)}) = X^2(\mathbf{n}^{\text{rep},(k)})$ (and/or other statistics of interest).

Step 5: Estimate the posterior predictive p value for a test statistic by the proportion of $T(\mathbf{n}^{\text{rep},(k)})$ which are greater than, or equal to $T(\mathbf{n})$:

$$\hat{p}_{test} = K^{-1} \sum_{k=1}^{K} I(T(\mathbf{n}^{\text{rep},(k)}) \geqslant T(\mathbf{n}^{\text{obs}})). \tag{2.13}$$

If $\hat{p}_{test}$ is less than a predefined value (for instance, .05) we conclude that the model does not fit the data properly.

PPCs are generally used to check whether specific aspects of the observed data are correctly picked up by the model (Gelman, Carlin, Stern, & Rubin, 2004). The BVR statistic is a good example of this, as it indicates one specific aspect of the model, rather than GoF at the aggregate

level. However, whether the X^2, G^2 or the BVR are suitable for use as test statistics in the PPC has yet to be determined.

An issue with the model-based PPC, which also holds for the parametric bootstrap, is that ML estimates have to be obtained for each of the replicated data sets. This makes both procedures rather time consuming, because the model has to be estimated for each replicated data set.

2.3.4 Parameter-based PPC

The added value of the parameter-based PPC (Gelman et al., 1996) over the model-based PPC and parametric bootstrap is that it not only incorporates uncertainty about the model parameters, but it also eliminates the need for model estimation for each replicated data set because we can define discrepancies $D(\mathbf{y};\boldsymbol{\theta})$ which not only depend the data $\mathbf{n}$ but also on the model parameters $\boldsymbol{\theta}$. This makes the parameter-based PPC computationally much faster than the other resampling methods.

Using the index k for a specific draw for the parameters obtained through the data augmentation algorithm, the parameter-based PPC proceeds as follows:

Algorithm 2.3: Parameter-based PPC in LC Analysis

Step 1: Obtain K draws $\boldsymbol{\theta}^{(k)}$ from the posterior distribution for the C-class model:

$$\boldsymbol{\theta}^{(k)} \sim p(\boldsymbol{\theta}|\mathbf{n}^{\text{obs}}), k = 1, \ldots, K.$$

Step 2: Calculate the estimated pattern probabilities $\widehat{P}(\mathbf{y}_s)^{(k)}$ from $\boldsymbol{\theta}^{(k)}$. Draw K random samples of size N from a multinomial distribution

with parameters $\widehat{P}(\mathbf{y}_s)^{(k)}$

$$\mathbf{n}^{\text{rep},(k)} \sim \text{Multin}(N, \widehat{P}(\mathbf{y}_1)^{(k)}, \ldots, \widehat{P}(\mathbf{y}_S)^{(k)})$$

Step 3: Calculate, for each data set $\mathbf{n}^{\text{rep},(k)}$ the realised discrepancies $D(\mathbf{n}^{\text{obs}}, \boldsymbol{\theta}^{(k)})$ and replicated discrepancies $D_{\text{rep}}^{(k)}$. For instance, when using the Pearson X^2:

$$D(\mathbf{n}^{\text{obs}}, \boldsymbol{\theta}^{(k)}) = X^2(\mathbf{n}^{\text{obs}}, \boldsymbol{\theta}^{(k)}) = \sum_{s=1}^{S} \frac{(n_s - e_s^{(k)})^2}{e_s^{(k)}} \tag{2.14}$$

and

$$D_{\text{rep}}^{(k)} = X^2(\mathbf{n}^{\text{rep},(k)}, \boldsymbol{\theta}^{(k)}) = \sum_{s=1}^{S} \frac{(n_s^{(k)} - e_s^{(k)})^2}{e_s^{(k)}}, \tag{2.15}$$

where the expected frequencies $e_s^{(k)} = NP(\boldsymbol{y}_s|\boldsymbol{\theta}^{(k)})$ (see Equation 2.1). The $n_s^{(k)}$ are the pattern frequencies in the replicated data set $\mathbf{n}^{\text{rep},(k)}$.

Step 4: Estimate the posterior predictive p value for a discrepancy by the proportion of replications for which $D_{\text{rep}}^{(k)}$ is greater than or equal to $D(\mathbf{n}^{\text{obs}}, \boldsymbol{\theta}^{(k)})$:

$$\hat{p}_{disc} = K^{-1} \sum_{k=1}^{K} I(D_{\text{rep}}^{(k)} \geqslant D(\mathbf{n}^{\text{obs}}, \boldsymbol{\theta}^{(k)})), \tag{2.16}$$

(where the indicator function I equals 1 if $D_{\text{rep}}^{(k)} \geqslant D(\mathbf{n}^{\text{obs}}, \boldsymbol{\theta}^{(k)})$ and 0 otherwise). If $\hat{p}_{disc}$ is close to 0 or 1, depending on what discrepancy is used, we conclude that the model does not fit the data properly (Gelman et al., 1996, 2004).

Note that Steps 1 and 2 for the parameter-based PPC are exactly the same as Steps 2 and 3 for the model-based PPC. But rather than comparing

replicated statistics to a single observed value based on the ML estimates, the parameter-based PPC compares K *pairs* of discrepancies; that is, K *realised* discrepancies, $D(\mathbf{n}^{\text{obs}}, \boldsymbol{\theta}^{(k)})$, with K *predictive* discrepancies, $D_{\text{rep}}^{(k)}$.

It is important to note that p_{disc}-values are different from the other p values in the sense that their distribution under the null-hypothesis is generally non-uniform (Meng, 1994) Rather, its distribution tends to be peaked around .5 (Robins, van der Vaart, & Ventura, 2000). Because of this, the parameter-based PPC will usually provide more conservative results and have lower power to reject a false model (Gelman, 2013).

2.4 Simulation Study

The quality of bootstrap and PPC p values for global and local GoF testing in LC analysis was investigated using two Monte Carlo studies. The first study evaluated the type I error rates. The second study investigated the power of the different methods and statistics. In both studies, p values were then obtained by either comparing the statistics to

1. a χ^2 distribution with given df,
2. the empirical distribution from the parametric bootstrap,
3. the empirical distribution from the model-based PPC, or
4. the empirical distributions from the parameter-based PPC.

We used the software package R 2.15.1 (R Core Team, 2012) to generate data sets, to perform the parameter-based PPC, and to collect the results. For ML estimation, asymptotic p value calculation, and the parametric bootstrap, we used LatentGOLD 5.0 (Vermunt & Magidson, 2013). The MCMC algorithm for the Bayesian LC analysis was implemented in

a routine written in C. We used a burn-in of 1000 iterations[1] and subsequently intervals of 10 iterations of the data augmentation algorithm between draws of $\boldsymbol{\theta}^{(k)}$.

2.4.1 Study 1. Type I errors

Design

To check type I error rates of the different p values, we fully crossed the following design factors:

- Sample size $N = 100$, 1000, or 5000.
- Number of LCs $C = 2$, or 3.
- Number of dichotomous variables $J = 6$, or 10.
- Conditional response probabilities $\pi_{1j1} = \pi_{2j2} = .7$, .8, or .9, for all j, and $\pi_{rj3} = \pi_{rj1}$, for $j = 1, \ldots, J/2$ and $\pi_{rj3} = \pi_{rj2}$ for $j = J/2 + 1, \ldots, J$

Table 2.1 provides the population parameters for each LC when $\pi_{1j1} = .8$. Additionally, we analysed conditions with J=6 trichotomous variables ($R_j = 3$ for all j) and sample sizes $N = 100$, 1000, or 5000. The population parameters are shown also in Table 2.1. In all conditions we generated 2000 data sets. Each data set was analysed using a LC model in which the number of classes was equal to the number of classes in the population model (i.e., the null-hypothesis was true). The parametric bootstrap was performed with $B = 500$ replications conditional on $\hat{\boldsymbol{\theta}}$. The model-based PPC and parameter-based PPC were performed based on $K = 500$ replications/draws.

[1]Inspection of the parameter estimates indicated that a burn-in of 1000 iterations was sufficient for our models, providing estimates comparable to the population parameters.

Table 2.1: Example of Population Parameters for the LCs, c, for conditions with $J = 6$ variables.

	$R_j = 2$					$R_j = 3$			
	$c = 1$	$c = 2$	$c = 3$	$c = 4$		$c = 1$	$c = 2$	$c = 3$	$c = 4$
π_c	.25	.25	.25	.25	π_c	.25	.25	.25	.25
π_{11c}	.8	.2	.8	.2	π_{11c}	.7	.1	.7	.1
π_{12c}	.8	.2	.8	.2	π_{21c}	.2	.2	.2	.2
π_{13c}	.8	.2	.8	.2	π_{12c}	.7	.1	.7	.1
π_{14c}	.8	.2	.2	.8	π_{22c}	.2	.2	.2	.2
π_{15c}	.8	.2	.2	.8	π_{13c}	.7	.1	.7	.1
π_{16c}	.8	.2	.2	.8	π_{23c}	.2	.2	.2	.2
					π_{14c}	.7	.1	.1	.7
					π_{24c}	.2	.2	.2	.2
					π_{15c}	.7	.1	.1	.7
					π_{25c}	.2	.2	.2	.2
					π_{16c}	.7	.1	.1	.7
					π_{26c}	.2	.2	.2	.2

Note. The conditional response probabilities for each class remain the same across the conditions. The class proportions are specified as $\pi_c = 1/C$ and thus vary across conditions.

We chose the simulation conditions such that the parameter values influence the level of sparseness but are also practically relevant. The chosen sample sizes of 100, 1000 and 5000 correspond typically to small, medium and large data sets, respectively. The sample size influences the degree of sparseness in the contingency table: The fewer respondents, the sparser the contingency table becomes.

The number of variables (6 or 10) and number of response categories affects the degree of sparseness. The number of possible patterns (i.e., cells in the table) was either $2^6 = 64$, $3^6 = 729$ or $2^{10} = 1024$. Note that in the $J = 10$ variable conditions, sparseness may be a problem even with a sample size of 5000.

Conditional response probabilities of .7, .8, and .9, respectively, indicate

a weak, medium and strong associations of the variables with the LCs. Note that these probabilities also influence the degree of sparseness besides sample size and the number of variables. When the conditional response probabilities of a particular response to an variable gets closer to 1, the number of patterns decreases, leading to an increase in sparseness.

Increasing the number of classes, on the other hand, decreases the sparseness of the contingency table, since the response preferences of each class lead to different response patterns. However, because this decrease in sparseness comes with an increased model complexity, it will be interesting to see any trade-off between model complexity and sparseness in determining the fit of a LC model.

Under the null-hypothesis, p values should be uniformly distributed (Sackrowitz & Samuel-Cahn, 1999). This also means that (approximately) 5% of the p values should fall below .05. We will therefore investigate the performance of the methods by checking whether the proportion of the simulation data sets yielding a p value less than .05 is close to .05.

Results

Results from study 1 on type I error rates can be found in Tables 2.2 through 2.5 for the dichotomous conditions and in Table 2.6 for the trichotomous conditions. The tables for the dichotomous conditions are arranged such that the least sparse condition is located top-left, meaning that by going downward or to the right, sparseness increases. For each combination of condition, fit-statistic and type of p value, we provide the proportion of simulations in which the obtained p value was less than .05. Due to expected fluctuations in 2000 replications per condition, we expect 99% of the p values to lie within the "expected interval" $.05 \pm 2.58\sqrt{.05(1-.05)/2000}$ (i.e., between 0.037 and 0.063). In the traditional context of null-hypothesis

testing this interval would signify close-to-nominal type I error rates. Proportions outside the interval may indicate problems with a given method, statistic, or combination of both and these proportions are underlined in the table. Note that for the BVR statistic, the asymptotic p values are based on a χ^2_1 distribution for the dichotomous and χ^2_4 for the trichotomous conditions, even though it has been shown to be incorrect. We include them to assess the practical implications of this common usage. No asymptotic p values are provided for the $TBVR$ and DI.

The standard GoF chi-squared statistics

Tables 2.2 and 2.3 provide the simulation results for the standard chi-squared GoF statistics for the dichotomous two-class and three-class conditions, respectively.

As expected, the asymptotic p values only provided close to nominal type I error rates for the situations where sparseness was not an issue. For $J = 6$ variables and $N = 5000$ or $N = 1000$ observations, the asymptotic p values may be useful, except when using the G^2. Asymptotic p values for G^2 only reached close-to-nominal type I error rates when there were 5000 observations.

The bootstrap and model-based PPC did considerably better than the asymptotic p values and performed comparably well, where serious problems only occurred in the most sparse condition of $J = 10$ and $N = 100$. The differences between parametric bootstrap and model-based PPC are generally small and mostly involve the G^2 statistic. For the G^2, the model-based PPC provides more conservative results than the parametric bootstrap in the $J = 10$ conditions for $N = 1000$ and $N = 100$.

Looking at the parameter-based PPC, we see that the proportions of p values less than .05 lie in the expected interval only in the $\pi_{1j1} = .7$,

$J = 6$, and $N = 5000$ or 1000 conditions. For the G^2 this also holds for the $\pi_{1j1} = .8$ and $\pi_{1j1} = .9$ conditions when $N = 5000$ and for the X^2 when $\pi_{1j1} = .9$ and $N = 5000$. For all other conditions the proportion of p values less than .05 was (much) less than .05, confirming the non-uniformity of the p_{disc}-value.

For the trichotomous conditions, the results found in Table 2.6 make it clear that the parametric bootstrap provides close-to-nominal type I-error rates in nearly all conditions and for all global fit statistics. Only in the $N = 100$ conditions were the type I error rates outside of the expected interval. The model-based PPC was overall a bit more conservative, even in the least sparse case. The parameter-based PPC was much too conservative in practically all conditions. Again it is shown that asymptotic p values are very unreliable, unless when used for the X^2 statistic in the $N = 5000$ conditions.

The results for the two- and three-class model are similar, albeit that the PPCs tend to get more conservative when model complexity increases. This effect is especially noticeable in the trichotomous conditions.

Statistics without a known asymptotic distribution

Tables 2.4 and 2.5 provide the simulation results for the BVR, $TBVR$, and DI for the dichotomous two-class and three-class conditions, respectively. These are all measures for which asymptotic p values are not available.

First of all, it can be observed that using the χ^2_1 distribution as the asymptotic reference distribution for the BVR is inadequate. The highest type I error rate was .0110, but generally these were much smaller still.

The parametric bootstrap generally works very well for the BVR, $TBVR$, and DI, with most proportions inside or very close to the expected interval, although it seems to work less well for the DI in the most extreme

sparseness condition. The model-based PPC had more proportions outside the expected interval, which generally resulted in somewhat more conservative conclusions. Overall, resampling techniques seem to work well when there is no reference distribution available.

For the parameter-based PPC, only in 1 condition, for the DI, a proportion of p values was found inside the expected interval. It can be seen that here, too, the p_{disc}-values are not uniformly distributed.

For the trichotomous conditions, the results found in Table 2.6 again show that the parametric bootstrap works very well when applied to the BVR, $TBVR$ and DI. It was only too conservative in the sparsest case of $N = 100$ and $C = 2$ in combination with the DI. The PPCs were too conservative, except for the PPC using the local fit measures as fit statistics in non-sparse conditions with two LCs. The BVR clearly did not follow a χ^2_4 distribution as the type I error rate for the p_{asymp} was 0 in all conditions.

The results for the two- and three-class model are similar, but again the PPCs become more conservative when model complexity increases.

Table 2.2: Type I Error Rates (the Proportion of p Values which were Less Than $\alpha = .05$) for the Global Fit Statistics based on 2000 MC Simulation Replications for the Conditions with 2 LCs.

		J=6					J=10				
		π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}	π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	G^2	.7	.056	.056	.058	.035	.7	.623	.053	.055	.054
		.8	.048	.044	.045	.026	.8	.591	.046	.047	.042
		.9	.063	.044	.046	.027	.9	.000	.054	.045	.043
	X^2	.7	.055	.057	.057	.031	.7	.057	.059	.057	.052
		.8	.043	.044	.043	.027	.8	.062	.049	.049	.027
		.9	.047	.048	.045	.023	.9	.252	.065	.065	.037
	CR	.7	.057	.057	.059	.031	.7	.055	.058	.056	.054
		.8	.043	.044	.045	.027	.8	.012	.050	.049	.031
		.9	.044	.048	.048	.024	.9	.000	.064	.064	.035
N=1000	G^2	.7	.052	.046	.042	.025	.7	.509	.051	.055	.048
		.8	.082	.059	.057	.032	.8	.000	.048	.033	.030
		.9	.098	.054	.056	.028	.9	.000	.038	.017	.017
	X^2	.7	.041	.044	.043	.028	.7	.126	.056	.054	.043
		.8	.055	.058	.057	.027	.8	.185	.058	.057	.016
		.9	.064	.051	.051	.021	.9	.357	.038	.038	.023
	CR	.7	.041	.042	.044	.026	.7	.005	.055	.059	.045
		.8	.051	.056	.053	.026	.8	.000	.059	.057	.015
		.9	.039	.053	.053	.021	.9	.000	.040	.040	.019
N=100	G^2	.7	.167	.072	.082	.033	.7	1.000	.096	.044	.022
		.8	.035	.069	.059	.021	.8	1.000	.026	.002	.002
		.9	.000	.073	.025	.011	.9	.905	.051	.000	.000
	X^2	.7	.041	.051	.054	.029	.7	1.000	.033	.021	.037
		.8	.042	.041	.044	.010	.8	1.000	.016	.016	.003
		.9	.140	.031	.048	.001	.9	1.000	.061	.076	.008
	CR	.7	.032	.060	.065	.032	.7	1.000	.118	.118	.034
		.8	.017	.050	.052	.015	.8	1.000	.054	.045	.001
		.9	.011	.054	.055	.003	.9	.999	.074	.074	.005

Table 2.3: Type I Error Rates (the Proportion of p Values which were Less Than $\alpha = .05$) for the Global Fit Statistics based on 2000 MC Simulation Replications for the Conditions with 3 LCs.

		J=6						J=10				
		π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}		π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	G^2	.7	.056	.058	.057	.023	G^2	.7	.702	.053	.052	.049
		.8	.044	.042	.044	.016		.8	.490	.052	.054	.051
		.9	.056	.043	.043	.017		.9	.000	.055	.050	.044
	X^2	.7	.055	.054	.057	.022	X^2	.7	.054	.059	.053	.048
		.8	.041	.044	.045	.014		.8	.060	.048	.048	.028
		.9	.040	.045	.041	.012		.9	.175	.057	.052	.029
	CR	.7	.053	.055	.057	.021	CR	.7	.051	.057	.056	.052
		.8	.044	.046	.046	.015		.8	.015	.049	.052	.035
		.9	.040	.045	.042	.012		.9	.001	.057	.057	.028
N=1000	G^2	.7	.068	.059	.055	.021	G^2	.7	.128	.067	.064	.055
		.8	.067	.045	.044	.015		.8	.000	.057	.035	.026
		.9	.097	.055	.056	.020		.9	.000	.047	.024	.019
	X^2	.7	.054	.060	.057	.019	X^2	.7	.133	.057	.055	.044
		.8	.043	.044	.043	.014		.8	.188	.058	.052	.012
		.9	.059	.052	.054	.014		.9	.319	.041	.042	.021
	CR	.7	.053	.061	.056	.020	CR	.7	.001	.065	.068	.048
		.8	.042	.045	.044	.014		.8	.000	.059	.057	.012
		.9	.048	.054	.055	.017		.9	.000	.048	.043	.015
N=100	G^2	.7	.097	.055	.056	.020	G^2	.7	1.000	.252	.028	.008
		.8	.029	.053	.063	.011		.8	1.000	.113	.001	.000
		.9	.003	.072	.034	.012		.9	1.000	.089	.000	.000
	X^2	.7	.059	.052	.054	.014	X^2	.7	1.000	.019	.024	.013
		.8	.029	.031	.061	.005		.8	1.000	.031	.052	.000
		.9	.083	.046	.078	.002		.9	1.000	.055	.076	.002
	CR	.7	.048	.054	.055	.017	CR	.7	1.000	.134	.118	.010
		.8	.008	.040	.073	.007		.8	1.000	.105	.087	.000
		.9	.015	.060	.073	.006		.9	1.000	.078	.071	.000

Table 2.4: Type I Error Rates (the Proportion of p Values which were Less Than $\alpha = .05$) for the BVR, the total BVR and the DI based on 2000 MC Simulation Replications for the Conditions with 2 LCs.

		J=6						J=10				
		π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}		π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	BVR	.7	.006	.065	.059	.000	BVR	.7	.011	.059	.057	.001
		.8	.002	.052	.050	.001		.8	.002	.060	.059	.001
		.9	.000	.045	.043	.001		.9	.000	.048	.049	.000
	$TBVR$	.7	NA	.051	.052	.000	$TBVR$	.7	NA	.051	.049	.001
		.8	NA	.053	.052	.001		.8	NA	.043	.043	.001
		.9	NA	.047	.048	.001		.9	NA	.049	.047	.000
	DI	.7	NA	.057	.055	.015	DI	.7	NA	.047	.044	.038
		.8	NA	.049	.050	.006		.8	NA	.049	.042	.028
		.9	NA	.052	.053	.001		.9	NA	.047	.039	.003
N=1000	BVR	.7	.002	.043	.039	.001	BVR	.7	.010	.049	.046	.001
		.8	.001	.061	.059	.000		.8	.003	.048	.044	.000
		.9	.000	.061	.061	.001		.9	.000	.048	.046	.000
	$TBVR$	.7	NA	.046	.042	.001	$TBVR$	.7	NA	.055	.058	.000
		.8	NA	.051	.045	.000		.8	NA	.047	.042	.000
		.9	NA	.051	.044	.001		.9	NA	.045	.034	.000
	DI	.7	NA	.053	.052	.016	DI	.7	NA	.058	.057	.041
		.8	NA	.053	.052	.006		.8	NA	.031	.009	.016
		.9	NA	.060	.051	.001		.9	NA	.046	.021	.004
N=100	BVR	.7	.007	.038	.037	.001	BVR	.7	.010	.042	.040	.001
		.8	.001	.052	.039	.000		.8	.004	.053	.044	.000
		.9	.000	.047	.036	.001		.9	.001	.059	.045	.000
	$TBVR$	.7	NA	.033	.033	.001	$TBVR$	.7	NA	.047	.031	.000
		.8	NA	.042	.021	.001		.8	NA	.053	.025	.000
		.9	NA	.048	.018	.001		.9	NA	.044	.007	.000
	DI	.7	NA	.071	.072	.018	DI	.7	NA	.045	.010	.015
		.8	NA	.062	.033	.003		.8	NA	.017	.000	.008
		.9	NA	.046	.007	.002		.9	NA	.027	.000	.006

Table 2.5: Type I Error Rates (the Proportion of p Values which were Less Than $\alpha = .05$) for the BVR, the total BVR, and the DI based on 2000 MC Simulation Replications for the Conditions with 3 LCs.

		J=6						J=10				
		π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}		π_{111}	p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	BVR	.7	.000	.043	.044	.001	BVR	.7	.005	.053	.051	.000
		.8	.000	.058	.057	.000		.8	.001	.052	.051	.000
		.9	.000	.056	.056	.000		.9	.001	.051	.053	.000
	$TBVR$	.7	NA	.052	.051	.001	$TBVR$	.7	NA	.054	.051	.000
		.8	NA	.061	.058	.000		.8	NA	.047	.045	.000
		.9	NA	.047	.046	.000		.9	NA	.058	.053	.000
	DI	.7	NA	.051	.052	.007	DI	.7	NA	.053	.049	.035
		.8	NA	.050	.050	.002		.8	NA	.049	.045	.023
		.9	NA	.043	.036	.000		.9	NA	.055	.048	.004
N=1000	BVR	.7	.000	.034	.023	.000	BVR	.7	.004	.042	.039	.000
		.8	.000	.047	.038	.000		.8	.001	.051	.050	.000
		.9	.000	.046	.037	.000		.9	.000	.052	.048	.000
	$TBVR$	.7	NA	.043	.033	.000	$TBVR$	.7	NA	.052	.048	.000
		.8	NA	.043	.033	.000		.8	NA	.054	.046	.000
		.9	NA	.045	.037	.000		.9	NA	.042	.038	.000
	DI	.7	NA	.049	.049	.007	DI	.7	NA	.058	.043	.035
		.8	NA	.037	.034	.002		.8	NA	.050	.029	.019
		.9	NA	.048	.044	.000		.9	NA	.046	.014	.006
N=100	BVR	.7	.000	.046	.037	.000	BVR	.7	.012	.044	.046	.000
		.8	.003	.016	.029	.000		.8	.003	.045	.037	.000
		.9	.001	.033	.030	.000		.9	.001	.043	.037	.000
	$TBVR$	.7	NA	.045	.037	.000	$TBVR$	.7	NA	.051	.048	.000
		.8	NA	.018	.025	.000		.8	NA	.040	.021	.000
		.9	NA	.036	.039	.000		.9	NA	.024	.012	.000
	DI	.7	NA	.048	.044	.000	DI	.7	NA	.135	.007	.008
		.8	NA	.054	.041	.003		.8	NA	.071	.000	.004
		.9	NA	.054	.005	.004		.9	NA	.051	.000	.003

Table 2.6: Type I Error Rates (the Proportion of p Values which were Less Than $\alpha = .05$) based on 2000 MC Simulation Replications for the Trichotomous Conditions, where $J = 6$ and $\boldsymbol{\pi}_{r11} = \{.7, .2, .1\}$

		$C = 2$					$C = 3$			
		p_{asymp}	p_{boot}	p_{test}	p_{disc}		p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	G^2	.585	.049	.021	.048	G^2	.516	.057	.009	.049
	X^2	.054	.044	.023	.033	X^2	.059	.052	.010	.029
	CR	.027	.049	.020	.038	CR	.033	.055	.011	.038
	BVR	.000	.046	.045	.004	BVR	.000	.055	.035	.001
	$TBVR$	NA	.051	.049	.000	$TBVR$	NA	.046	.033	.000
	DI	NA	.040	.010	.021	DI	NA	.059	.006	.024
N=1000	G^2	.000	.055	.015	.039	G^2	.000	.056	.005	.035
	X^2	.087	.053	.031	.024	X^2	.083	.048	.017	.020
	CR	.000	.058	.029	.020	CR	.001	.059	.011	.021
	BVR	.000	.046	.045	.002	BVR	.000	.043	.022	.004
	$TBVR$	NA	.054	.050	.001	$TBVR$	NA	.051	.026	.000
	DI	NA	.045	.007	.022	DI	NA	.047	.002	.017
N=100	G^2	1.000	.038	.000	.001	G^2	1.000	.075	.000	.001
	X^2	1.000	.020	.008	.002	X^2	1.000	.022	.021	.001
	CR	1.000	.070	.017	.005	CR	1.000	.079	.023	.000
	BVR	.000	.040	.032	.004	BVR	.001	.046	.026	.000
	$TBVR$	NA	.036	.018	.000	$TBVR$	NA	.040	.010	.000
	DI	NA	.021	.000	.006	DI	NA	.051	.000	.004

2.4.2 Study 2. Power Analysis

Design

After evaluating type I error rates, we also investigated power of the different p values. Power is the probability of rejecting a model when it is indeed false. To do this, we estimated a two-class model on data sets generated under a three-class population, and estimated a three-class model on data sets generated under a four-class population. Population parameters for these conditions were $\pi_{1j1} = .8$ with $J = 6$, or 10 variables in the dichotomous cases, and $\pi_{1j1} = .7$ with $J = 6$ variables in the trichotomous cases (cf. Table 2.1). For each condition 2000 data sets were generated and analysed.

Results

Results of the power analysis for the dichotomous conditions can be found in Tables 2.7 and 2.8 and for the trichotomous conditions in Table 2.9. Power of .8 or greater is generally regarded to be acceptable, and higher values are better. It is immediately clear that the power to detect that a model has too few LCs is very high in medium ($N = 1000$) to large ($N = 5000$) data sets, as most of the power values are 1.0. For small data sets ($N = 100$) the power was around .2, though it is noteworthy that the $TBVR$, when used in the parametric bootstrap or as statistic in the PPC, has high power even in the sparsest condition when $C = 2$. Also, the power to detect misfit using the $TBVR$ increases as the number of variables increases.

In order to draw conclusions about the usefulness of the methods and statistics, we need to combine the results of Study 1 and 2. For example, when a statistic has high power but also has large type I error rates (larger

than the chosen level of significance α), the statistic will lead to too liberal results and general use is not recommended. In such cases we would have a high chance of rejecting a model, regardless of whether the model is actually true or false. For the p_{boot} and p_{test}, power was high and type I error rates were very accurate in most conditions as well. The p_{test} is overall somewhat more conservative. The p_{disc} had very low type I error rates but still had high power to detect the misspecification of the models in our simulation by means of the global chi-squared statistics. The p_{asymp} also showed high power, but also had very high type I error rates when sparseness became an issue (e.g. when $J = 10$). When assessing LC model fit using any particular statistic, we advise researchers to use either the parametric bootstrap or PPC using fit statistics. When tables are sparse due to small sample sizes, researchers should resort to local fit statistics, which may be tailored to the research question at hand.

Table 2.7: Power (the Proportion of p values which were Less Than $\alpha = .05$) to Indicate Model Misfit when a Model with C Classes is Estimated on Data Generated under Population with $C + 1$ LCs. Conditions with $J = 6$ Dichotomous variables where $\boldsymbol{\pi}_{r11} = \{.8, .2\}$. Results are based on 2000 MC Simulations.

		C = 2					C = 3			
		p_{asymp}	p_{boot}	p_{test}	p_{disc}		p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	G^2	1.000	1.000	1.000	1.000	G^2	1.000	1.000	1.000	1.000
	X^2	1.000	1.000	1.000	1.000	X^2	1.000	1.000	1.000	1.000
	CR	1.000	1.000	1.000	1.000	CR	1.000	1.000	1.000	1.000
	BVR	1.000	1.000	1.000	.973	BVR	1.000	1.000	1.000	.966
	$TBVR$	NA	1.000	1.000	1.000	$TBVR$	NA	1.000	1.000	1.000
	DI	NA	1.000	1.000	1.000	DI	NA	1.000	1.000	1.000
N=1000	G^2	1.000	1.000	1.000	1.000	G^2	1.000	1.000	1.000	1.000
	X^2	1.000	1.000	1.000	1.000	X^2	1.000	1.000	1.000	1.000
	CR	1.000	1.000	1.000	1.000	CR	1.000	1.000	1.000	1.000
	BVR	.711	.922	.914	.513	BVR	.515	.980	.974	.088
	$TBVR$	NA	1.000	1.000	1.000	$TBVR$	NA	1.000	1.000	.0110
	DI	NA	1.000	1.000	1.000	DI	NA	1.000	1.000	1.000
N=100	G^2	.456	.524	.497	.353	G^2	.156	.163	.160	.056
	X^2	.413	.409	.428	.257	X^2	.065	.093	.132	.023
	CR	.308	.488	.496	.313	CR	.045	.125	.152	.036
	BVR	.228	.322	.321	.048	BVR	.014	.110	.130	.000
	$TBVR$	NA	.717	.690	.002	$TBVR$	NA	.133	.118	.000
	DI	NA	.546	.492	.310	DI	NA	.217	.181	.039

Table 2.8: Power (the Proportion of p values which were Less Than $\alpha = .05$) to Indicate Model Misfit when a Model with C Classes is Estimated on Data Generated under Population with $C + 1$ LCs. Conditions with $J = 10$ Dichotomous variables where $\boldsymbol{\pi}_{r11} = \{.8, .2\}$. Results are based on 2000 MC Simulations.

		C = 2					C = 3			
		p_{asymp}	p_{boot}	p_{test}	p_{disc}		p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	G^2	1.000	1.000	1.000	1.000	G^2	1.000	1.000	1.000	1.000
	X^2	1.000	1.000	1.000	1.000	X^2	1.000	1.000	1.000	1.000
	CR	1.000	1.000	1.000	1.000	CR	1.000	1.000	1.000	1.000
	BVR	.973	.993	.992	.861	BVR	.959	1.000	1.000	.733
	$TBVR$	NA	1.000	1.000	1.000	$TBVR$	NA	1.000	1.000	1.000
	DI	NA	1.000	1.000	1.000	DI	NA	1.000	1.000	1.000
N=1000	G^2	1.000	1.000	1.000	1.000	G^2	1.000	1.000	1.000	1.000
	X^2	1.000	1.000	1.000	1.000	X^2	1.000	1.000	1.000	1.000
	CR	1.000	1.000	1.000	1.000	CR	1.000	1.000	1.000	1.000
	BVR	.614	.761	.759	.501	BVR	.514	.976	.954	.453
	$TBVR$	NA	1.000	1.000	1.000	$TBVR$	NA	1.000	1.000	.817
	DI	NA	1.000	1.000	1.000	DI	NA	1.000	.999	1.000
N=100	G^2	1.000	.485	.124	.100	G^2	1.000	.355	.010	.006
	X^2	1.000	.116	.126	.056	X^2	1.000	.012	.018	.002
	CR	1.000	.286	.258	.055	CR	1.000	.117	.069	.002
	BVR	.237	.316	.320	.073	BVR	.047	.167	.162	.003
	$TBVR$	NA	.967	.957	.002	$TBVR$	NA	.567	.463	.001
	DI	NA	.290	.092	.179	DI	NA	.262	.003	.029

Table 2.9: Power (the Proportion of p values which were Less Than $\alpha = .05$) to Indicate Model Misfit when a Model with C Classes is Estimated on Data Generated under Population with $C + 1$ LCs. Conditions with $J = 6$ Trichotomous variables where $\boldsymbol{\pi}_{r11} = \{.7, .2, .1\}$. Results are based on 2000 MC Simulations.

		C = 2					C = 3			
		p_{asymp}	p_{boot}	p_{test}	p_{disc}		p_{asymp}	p_{boot}	p_{test}	p_{disc}
N=5000	G^2	1.000	1.000	1.000	1.000	G^2	1.000	1.000	1.000	1.000
	X^2	1.000	1.000	1.000	1.000	X^2	1.000	1.000	1.000	1.000
	CR	1.000	1.000	1.000	1.000	CR	1.000	1.000	1.000	1.000
	BVR	.838	1.000	1.000	.870	BVR	.967	1.000	1.000	.992
	$TBVR$	NA	1.000	1.000	1.000	$TBVR$	NA	1.000	1.000	1.000
	DI	NA	1.000	1.000	1.000	DI	NA	1.000	1.000	1.000
N=1000	G^2	1.000	1.000	1.000	1.000	G^2	1.000	1.000	.996	1.000
	X^2	1.000	1.000	1.000	1.000	X^2	.998	.996	.964	.984
	CR	1.000	1.000	1.000	1.000	CR	.980	1.000	.991	.996
	BVR	.600	.840	.813	.509	BVR	.514	.976	.954	.453
	$TBVR$	NA	1.000	1.000	1.000	$TBVR$	NA	1.000	1.000	.817
	DI	NA	1.000	1.000	1.000	DI	NA	1.000	.999	1.000
N=100	G^2	1.000	.338	.013	.062	G^2	1.000	.221	.001	.006
	X^2	1.000	.071	.053	.014	X^2	1.000	.022	.017	.002
	CR	1.000	.215	.111	.018	CR	1.000	.096	.030	.002
	BVR	.140	.480	.463	.086	BVR	.001	.295	.224	.003
	$TBVR$	NA	.963	.932	.001	$TBVR$	NA	.614	.312	.000
	DI	NA	.283	.006	.147	DI	NA	.187	.000	.019

2.5 Empirical Data

We will illustrate the methods described in the chapter with a data set taken from Landis and Koch (1977) (see also, Holmquist, McMahan, and Williams (1968)). It contains information on 118 slides which were evaluated on the absence or presence of cervical cancer by seven pathologists. So, we have a data set with 7 dichotomous variable and a sample size of 118. Only 20 of the possible $2^7 = 128$ response patterns were observed, indicating that we are dealing with a rather sparse contingency table. This sparse table has been used by various authors who proposed using bootstrap p values for global fit testing with G^2 (Agresti, 2002; Magidson & Vermunt, 2004; Vermunt & Magidson, 2005) Here, we will also look at other measures and consider both PPCs in addition to the bootstrap.

We estimated LC models with two or three LCs and assessed the GoF of these two models based on the X^2, G^2, CR, BVR, $TBVR$, and DI statistics. Results from these analyses can be found in Table 2.10.

Table 2.10: Fit Statistics and p values for the Cervical Cancer Data. Model with 2 or 3 LCs.

	Model with 2 LCs						Model with 3 LCs				
	Value	p_{asymp}	p_{boot}	p_{test}	p_{disc}		Value	p_{asymp}	p_{boot}	p_{test}	p_{disc}
G^2	64.163	1.000	.000	.012	.020	G^2	17.713	1.000	.500	.948	.662
X^2	90.564	.800	.028	.144	.332	X^2	21.120	1.000	.296	.870	.852
CR	74.851	.980	.006	.062	.212	CR	18.589	1.000	.360	.908	.828
$TBVR$	32.281	NA	.000	.000	.656	$TBVR$	8.328	NA	.026	.042	.586
DI	.268	NA	.000	.000	.010	DI	.117	NA	.146	.598	.298
BVR_{12}	1.736	.188	.028	.022	.314	BVR_{12}	.051	.822	.332	.442	.364
BVR_{13}	.387	.534	.090	.188	.838	BVR_{13}	.092	.762	.318	.370	.804
BVR_{14}	.273	.601	.196	.360	.668	BVR_{14}	.575	.448	.120	.124	.612
BVR_{15}	.146	.702	.456	.422	.410	BVR_{15}	.162	.687	.416	.290	.430
BVR_{16}	.209	.648	.362	.438	.628	BVR_{16}	.043	.836	.670	.668	.526
BVR_{17}	.024	.878	.682	.542	.442	BVR_{17}	.152	.697	.428	.338	.386
BVR_{23}	.017	.896	.824	.852	.648	BVR_{23}	.006	.939	.794	.838	.648
BVR_{24}	.577	.447	.190	.228	.736	BVR_{24}	.599	.439	.172	.172	.704
BVR_{25}	8.443	.004	.000	.000	.204	BVR_{25}	.036	.850	.290	.364	.336
BVR_{26}	.445	.505	.302	.332	.618	BVR_{26}	.477	.490	.256	.236	.548
BVR_{27}	5.205	.023	.000	.000	.304	BVR_{27}	.029	.866	.532	.442	.354
BVR_{34}	.895	.344	.256	.240	.782	BVR_{34}	.019	.890	.774	.726	.568
BVR_{35}	1.106	.293	.042	.056	.786	BVR_{35}	.134	.715	.114	.320	.788
BVR_{36}	1.316	.251	.160	.158	.772	BVR_{36}	.021	.886	.820	.758	.562
BVR_{37}	.138	.711	.098	.278	.832	BVR_{37}	.078	.780	.436	.376	.760
BVR_{45}	.043	.836	.814	.746	.586	BVR_{45}	.701	.403	.092	.112	.546
BVR_{46}	7.228	.007	.006	.002	.950	BVR_{46}	4.521	.033	.004	.004	.826
BVR_{47}	.099	.753	.236	.418	.622	BVR_{47}	.426	.514	.210	.194	.592
BVR_{56}	.589	.443	.248	.204	.612	BVR_{56}	.070	.792	.286	.550	.522
BVR_{57}	3.331	.068	.000	.000	.328	BVR_{57}	.101	.751	.356	.282	.372
BVR_{67}	.075	.785	.286	.520	.584	BVR_{67}	.038	.846	.664	.628	.526

Asymptotic p values are not appropriate here due to the sparseness of the contingency table. Based on the simulation results of sparse tables, we expect to see that, overall, the model-based PPC provides somewhat more conservative p values than the parametric bootstrap does.

Indeed, for the two-class model the parametric bootstrap provides p values of .028 for the X^2, .000 for the G^2 and .006 for the CR, indicating that the model is inadequate. The p_{test}-values were .144, .012 and .062, respectively, meaning only the G^2 statistic suggests lack of fit for the two-class model. The p_{disc} values were .332, .020 and .212 respectively. Here too, only the G^2 statistic indicated lack of fit.

Inspection of the bivariate residuals for the two-class model shows that some association remains between the variable pairs {2,5}, {2,7}, {5,7} and {4,6}. Asymptotic p values based on the χ^2_1 distribution indicate significant remaining associations, except perhaps for the BVR of variables 5 and 7 ($p_{asymp} = .058$). The parametric bootstrap and model-based PPC both indicate that these remaining associations are significantly different from 0. The parameter-based PPC did not provide p values close to zero. However, the most extreme p_{disc}-values are generally found for the largest BVR. For BVR_{46} the p value was .950, which also indicates misfit.

The parametric bootstrap and model-based PPC both indicate model misfit with regard to the $TBVR$ and DI, with p values of .000. The parameter-based PPC only indicated lack of fit for the DI and not for the $TBVR$.

For the three-class model, all methods indicate that the global fit of the model is adequate, based on the X^2, G^2 and CR and DI. As we expected from the simulation results, the p_{test}-values for these statistics were larger than those from the bootstrap.

Inspection of the bivariate residuals reveals that the association between

the variable pair {4,6} is not picked up by the three-class model (BVR_{46} = 4.521). The parametric bootstrap and model-based PPC statistics both indicate that this remaining association is significantly different from 0. The parameter-based PPC did not provide extreme p values here. This agrees with the simulation in which we virtually never saw p_{disc}-values for the BVR less than .05.

The parametric bootstrap and model-based PPC are able to pick up that there is remaining bivariate association through the $TBVR$ statistic, as both techniques provided small p values for this statistic.

In summary, the analyses show that a three-class model has adequate overall fit, but lacks in local fit, as indicated by the p values for BVR_{46} and for the $TBVR$. Also, the empirical data analysis was in agreement with our expectations from the simulation study that the model-based PPC yields somewhat more conservative p values than the parametric bootstrap does.

2.6 Discussion

To assess the fit of a LC model when contingency tables are sparse or when asymptotic reference distributions are not available, resampling techniques can be used to obtain empirical reference distributions for any goodness-of-fit statistics. In the current chapter we evaluated a number of statistics which are commonly used in the assessment of model fit, some of which are specific to LC models. We conducted a simulation study to investigate whether reliable p values could be obtained with the parametric bootstrap, the model-based PPC, and the parameter-based PPC.

The simulation study involved calculating different p values when analysing sparse and non-sparse contingency tables both for fit statistics that have no known asymptotic distribution, as well as for statistics for which the

asymptotic distributions do not hold in sparse situations. In agreement with previous studies we found that the use of asymptotic p values resulted in (severely) distorted type I error rates when contingency tables were sparse. Both the parametric bootstrap and model-based PPC performed much better in this regard than the asymptotic method. Von Davier (1997) showed that the likelihood-ratio G^2 is not suitable for use in the parametric bootstrap when contingency tables are sparse, as it will generally lead to too liberal conclusions. We have replicated this finding and have additionally shown that using the G^2 as a statistic in the PPC resulted in too conservative results. Because the G^2 had high type I error rates and high power, we cannot be sure what a small p_{asymp}-value indicates. The Pearson X^2 and CR however, did generally provide close-to-nominal type I error rates and had high power. These latter should therefore in many situations be preferred over the likelihood-ratio statistic G^2.

The DI statistic worked very well in combination with the parametric bootstrap and with the model-based PPC. The model-based PPC provides somewhat more conservative p values. Only in the most sparse condition did the parametric bootstrap show severe problems. The DI appears therefore be a good statistic to assess global model fit, even when contingency tables are sparse. However, when sample size is small ($N = 100$), it lacks power like the other global chi-squared statistics.

Sparseness has little effect on the BVR statistics, especially for dichotomous variables, as it only involves the second order marginals of the contingency tables. Therefore, it may be hypothesised that the use of asymptotics is justified. However, we have shown in line with Oberski et al. (2013) that the common distributional assumption for the BVR does not hold for LC models. Use of the χ^2_1-distribution produced too conservative results (i.e., low type I error rates). We would like to stress that the

poor results ascribed to the asymptotic p values for the BVR statistics are due to the choice for this reference distribution. Future research should indicate which, if any, asymptotic reference distribution should be used for the BVR in LC analysis.

For the BVR, both the parametric bootstrap and model-based PPC resulted in close-to-nominal type I error rates, even when the tables were very sparse. The latter method provided somewhat more conservative results than the former. The BVR statistic failed completely in combination with the parameter-based PPC.

The parametric bootstrap yielded close-to-nominal type I error rates when using the $TBVR$. In combination with the model-based PPC, the $TBVR$ resulted in somewhat below-nominal type I error rates. The power of the $TBVR$ was very high, however, and it shows that taking all bivariate associations into account provides very good information on model fit, even when tables are very sparse. Note that the findings of the BVR and $TBVR$, the latter being the sum of all BVRs, should not be seen as independent.

Our power study suggested that all methods and statistics are useful to detect misfit when the number of LCs is misspecified. When sample sizes become very small, however, the results have shown that we should resort to the local fit measures. Especially the $TBVR$ has very high power, since it is not greatly affected by sparseness and still uses information on all variable pairs to indicate whether misfit is present. Since no asymptotic distribution is known for this statistic, its use in the parametric bootstrap and as a statistic in the PPC will show to be of great value, even when data is sparse.

To illustrate our findings we analysed an empirical data set where, due to sparseness, the use of asymptotic p values was inadequate. We obtained

alternative p values by means of the parametric bootstrap, the model-based PPC and parameter-based PPC. In line with the results from the simulation study, we found that the model-based PPC provided somewhat more conservative results than the parametric bootstrap. Bootstrapping the global fit statistics strongly suggested that a two-class model did not fit the data adequately. However, when incorporating the uncertainty about the parameter estimates in the analysis, the model-based PPC did not provide very strong evidence to suggest model misfit. No disagreement was found between the parametric bootstrap and model-based PPC with regard to the BVR, $TBVR$ and DI statistics. In the better fitting three-class model, all methods indicated no lack of global fit. A nice result was that the parametric bootstrap and model-based PPC were well able to pick up violations in the local fit through the BVR and $TBVR$ statistics, even though the global fit measures indicated no problems.

Overall, the computationally less intensive parameter-based PPC provided more conservative results than the other resampling techniques. This was, to an extent, expected from the fact that the distribution of p_{disc} under the null-hypothesis is peaked around .5. A number of methods have been proposed to adjust the p_{disc} value so that it provides uniform p values (Bayarri & Berger, 2000; Hjort, Dahl, & Steinbakk, 2006; Robins et al., 2000). Future research should indicate whether extra computational burden of calibrating p_{disc}-values outweighs the benefits, compared to the properly working model-based PPC.

Given the established results, researchers should be weary of using asymptotic reference distributions when the sample sizes are not very large and/or when there are many variable, leading to a sparse contingency table. Resorting to lower-order statistics, like the BVR, and statistics which are specifically tailored to a certain application or research question, like the

DI, is good practice, even if their distributions are unknown. Though developing new statistics was not our aim here, many others can be conceived of. For dichotomous data, one could use a bivariate Pearson correlation to assess local dependencies. When interest lies in a specific second-order relationship, trivariate residuals could be used. If one response pattern is of particular interest, one could use the observed frequency of that pattern as a statistic. In each of these cases, resampling methods can provide reliable p values. Also, in the very sparse cases, using resampling techniques to assess combinations of the lower level associations, like the $TBVR$ proved to be very useful.

On a final note, this chapter addressed the question of assessing model *fit* and not model *comparison*. Interpretation of, for instance, information criteria like the AIC and BIC does not change when the sample size is small or when contingency tables are sparse.

Chapter 3

Posterior Calibration of Posterior Predictive p Values

Abstract

In order to accurately control the type I error rate (typically .05), a p value should be uniformly distributed under the null model. The posterior predictive p value (ppp), which is commonly used in Bayesian data analysis, generally does not satisfy this property. For example, there have been reports where the sampling distribution of the ppp under the null model was highly concentrated around .50. In this case, a ppp of .20 would indicate model misfit, but when comparing it with a significance level of .05, which is standard statistical practice, the null model would not be rejected. Therefore, the ppp has very little power to detect model misfit. A solution has been proposed in the literature, which involves calibrating the ppp using the prior distribution of the parameters under the null model. A disadvantage of this "prior-cppp" is that it is very sensitive to the prior of the model parameters. In this chapter, an alternative solution is proposed where the

This chapter is published as van Kollenburg, G.H., Mulder, J. & Vermunt, J.K. (2017). Posterior Calibration of Posterior Predictive p Values. *Psychological Methods, 22*(2), 382–396.

ppp is calibrated using the posterior under the null model. This "posterior-cppp" (i) can be used when prior information is absent, (ii) allows one to test any type of misfit by choosing an appropriate discrepancy measure, and (iii) has a uniform distribution under the null model. The methodology is applied in various testing problems such as testing independence of dichotomous variables, checking misfit of linear regression models in the presence of outliers, and assessing misfit in latent class analysis.

3.1 Introduction

A crucial step in a statistical analysis is to assess whether the employed statistical model fits the observed data. Different tools are available for this purpose. When one is interested in testing whether one statistical model better fits the data than another statistical model, *model comparison* tools are useful. Commonly used model comparison tools are the AIC (Akaike, 1973), the BIC (Schwarz, 1978), or the Bayes factor (Jeffreys, 1961; Kass & Raftery, 1995). These criteria penalize model complexity in the sense that a simple model with few parameters is generally preferred over a more complex model with many parameters if both models fit the data equally well (Myung, 2000; Mulder, 2014). On the other hand, when one is interested in testing whether one specific model fits the observed data, *model checking* tools are useful. Such model checks are often performed using Fisherian p values in a classical framework and using posterior predictive p values (ppp's) in a Bayesian framework (Meng, 1994; Berkhof et al., 2003; Choi, Hui, & Bell, 2010). Although these methods come from different paradigms, the p value and the ppp are applied in a similar fashion to assess misfit of the employed statistical model. If the p value is smaller than a pre-specified threshold value (typically .05), this indicates model misfit. If this is the case, an extension of the employed model may be

necessary to better fit the observed data. Because the p value and ppp are applied in a similar manner in practice, we shall also refer to this threshold value as the "significance level" of the Bayesian test to avoid additional terminology (despite the fact that the term "significance" is not commonly used in Bayesian statistics). We shall also borrow frequentist terminology when referring to the type I error rate, the type II error rate, and power as the probability of incorrectly rejecting a Bayesian null model that is true, the probability of not rejecting an incorrect null model, and the probability of correctly rejecting an incorrect null model, respectively. Note that such frequentist properties are of interest in objective Bayesian statistics (Berger, 2006).

In this chapter we shall focus on model checking in the Bayesian framework using the posterior predictive p value (ppp) (Gelman et al., 1996; Meng, 1994). The ppp has three useful properties. Firstly, it can straightforwardly be used for testing any type of model misfit; we only need to formulate a discrepancy measure which is able to detect the type of misfit of interest. Secondly, a ppp can easily be computed from MCMC output because the discrepancy is allowed to depend on the (sampled) unknown model parameters. Thirdly, the ppp can be computed using non-informative (or objective) improper priors. This third property is useful when prior information is unavailable or when a researcher does not want to include external information via the prior distribution when evaluating the model. Because of these useful properties, the ppp has been used in many different types of applications in psychology (Oravecz, Faust, Batchelder, & Levitis, 2015), as well as in other fields such as marketing (Choi et al., 2010), medicine (Chaves, Chakraborty, Benziger, & Tannenbaum, 2014), psychiatry (Berkhof et al., 2003) and sociology (Hoverd & Sibley, 2013).

A potential problem of the ppp, however, lies in its interpretation. In

order to reliably interpret a ppp, the type I error rate should be equal to the significance level. This equality only holds when the sampling distribution of the p value is uniform. For the ppp, however, the sampling distribution under the null model is typically not uniformly distributed, but instead it is concentrated around .5 (Meng, 1994), possibly with a lower bound that is larger than 0 (Robins et al., 2000). As a result, the type I error rate for the ppp is generally lower than the significance level. Consequently a badly fitting model may not be rejected as a result of very low statistical power of the ppp.

To resolve this issue and to accurately control the type I error probability, one can calibrate the ppp under the employed null model. For this purpose, Hjort et al. (2006) proposed calibrating the ppp using a proper informative prior, yielding what we will refer to as a prior-calibrated ppp (prior-cppp). The prior-cppp is uniformly distributed under the null model and the chosen prior, and therefore it potentially resolves the problem associated with the standard ppp.

A key property of the prior-cppp is however that the employed statistical model and the informative prior are simultaneously tested. It is therefore crucial to carefully formulate informative priors for the unknown model parameters based on one's substantive beliefs before observing the data. When prior information is weak or unavailable the prior-cppp is not recommendable. On the other hand when prior knowledge is available, the specification of the informative prior distribution for all model parameters can be a rather difficult and time consuming exercise (Berger, 2006; Hjort et al., 2006), which substantive researchers may prefer to avoid. Furthermore, researchers may only be interested in assessing model misfit of the employed statistical model and not in simultaneously testing the appropriateness of the informative prior. By taking these considerations into

account, the applicability of the prior-cppp may be limited.

In this chapter, a new type of calibrated ppp is proposed. Instead of calibrating the ppp under an informative prior, as in the prior-cppp, the ppp is calibrated under the posterior distribution of the unknown parameters of the employed null model. The resulting ppp will be referred to as the posterior-calibrated ppp (posterior-cppp). Unlike the prior-cppp, the posterior-cppp can be used when prior information is weak or when one is only interested in testing model misfit. Furthermore, the posterior-cppp has all the useful properties of the original ppp, with the additional advantage that it is uniformly distributed under the null model.

The remainder of the chapter is outlined as follows. In the next section we explain how to obtain the ppp, the prior-cppp, and the posterior-cppp. Subsequently, in Application I, the performance of the three posterior predictive checks is assessed for a simple test for independence in contingency tables by looking at type I error probabilities. In Application II, we apply the new methodology in a linear regression analysis to test whether the model adequately explains extreme observations. A simulation experiment is conducted to evaluate type I error rates and the power of this test. The method is applied in an empirical example to predict the quality of life of elderly people. In Application III, the methodology is applied in the context of latent class analysis. We investigate type I error rates and power when testing bivariate residuals and the number of latent classes. An empirical data set is used to illustrate the practical use and benefits of the posterior-cppp in testing for different sub-types of depression. The chapter ends with a discussion of the methods and results.

3.2 Posterior Predictive Checks

The posterior predictive check is a flexible and efficient tool to assess misfit of a Bayesian statistical model for the observed data. The general idea of a posterior predictive check is to assess systematic discrepancies between the observed data and (hypothetical) replicated data generated from the fitted model (Gelman et al., 2004). When there is a small discrepancy between the replicated data and the observed data, this suggests a good fit of the model. When there is a large discrepancy between the replicated data and the observed data, this suggests model misfit.

The procedure works as follows. First we have to specify a prior distribution for the unknown model parameters $\boldsymbol{\theta}$. The prior contains our knowledge or beliefs about the model parameters before observing the data. The prior will be denoted by $p(\boldsymbol{\theta})$. The model can be fitted to the observed data by deriving the posterior distribution of $\boldsymbol{\theta}$. The posterior is a combination of the information in the prior, $p(\boldsymbol{\theta})$, and the information in the observed data, $\mathbf{y}_{\text{obs}}$. The information about $\boldsymbol{\theta}$ in the observed data is formalised in the likelihood function of the model, which is denoted by $p(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta})$. Subsequently the posterior can be obtained using Bayes' theorem,

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) &= \frac{p(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})}{p(\mathbf{y}_{\text{obs}})} \\ &\propto p(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}), \end{aligned} \tag{3.1}$$

where $p(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}})$ denotes the posterior of the unknown parameters $\boldsymbol{\theta}$ given the observed data $\mathbf{y}_{\text{obs}}$. The posterior contains our knowledge about the model parameters after observing the data. In (3.1), the marginal likelihood $p(\mathbf{y}_{\text{obs}})$ does not depend on $\boldsymbol{\theta}$, and therefore it does not play a role when deriving the posterior. Due to the many possible specifications

of likelihood and prior, the posterior does not always belong to a known family of probability distributions. In such cases, the posterior is usually approximated by sampling posterior draws of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}})$ using an MCMC algorithm (for an extensive overview of MCMC algorithms, see Liang, Liu, & Carroll, 2011).

The posterior can be used to obtain estimates for the model parameters (such as posterior means, modes, or medians) and credibility intervals (the Bayesian counterpart of classical confidence intervals). Furthermore, the posterior can be used to draw a replicated data set, denoted by $\mathbf{y}_{\text{rep}}$. A replicated data set can be viewed as data that we could see tomorrow if the experiment that produced the observed data, $\mathbf{y}_{\text{obs}}$, were replicated with the same model and with the same value for $\boldsymbol{\theta}$ that produced the observed data today (Gelman et al., 2004). The posterior predictive distribution of $\mathbf{y}_{\text{rep}}$ is given by

$$p(\mathbf{y}_{\text{rep}}|\mathbf{y}_{\text{obs}}) = \int p(\mathbf{y}_{\text{rep}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}})d\boldsymbol{\theta}.$$

By looking at specific characteristics of the observed data and a replicated data set we can check whether both data sets were likely to be generated from the employed statistical model, similar as a classical test. If this is (not) the case, this suggests a good (bad) model fit to the observed data. This can be done using a so-called discrepancy measure, denoted by $D(\mathbf{y};\boldsymbol{\theta})$, which is a function of a data set $\mathbf{y}$ (which could be either the observed data set, $\mathbf{y}_{\text{obs}}$, or a replicated data set, $\mathbf{y}_{\text{rep}}$) and the unknown model parameters $\boldsymbol{\theta}$. For example, discrepancies can measure overall fit based on Pearson χ^2-type statistics (van Kollenburg, Mulder, & Vermunt, 2015), or specific aspects of the model such as adequately capturing extreme values (Gelman et al., 2004). Note that discrepancies can depend on

both the data and the model parameters (Meng, 1994), while classical fit statistics only depend on the data. Examples of discrepany measures will be provided in the next sections. Through the posterior predictive check we assess the probability – quantified by the ppp – that replicated data under the posterior are more extreme than the observed data (Gelman et al., 2004). The following algorithm describes how to obtain the ppp.

Algorithm 3.1: Computation of the ppp

Step 1: Specify a prior, $p(\boldsymbol{\theta})$, for the model parameters and choose a discrepancy measure, $D(\mathbf{y}; \boldsymbol{\theta})$.

Step 2: Obtain the posterior based on the prior and the likelihood of the observed data using (3.1).

Step 3: Obtain values for the chosen discrepancy measure for the observed data and replicated data sets based on random draws from the posterior:

3a: Draw a random value for the model parameters, denoted by $\boldsymbol{\theta}^{(k)}$, from the posterior:

$$\boldsymbol{\theta}^{(k)} \sim p(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}). \tag{3.2}$$

3b: Sample a replicate data set, $\mathbf{y}_{\text{rep}}^{(k)}$, given the posterior draw $\boldsymbol{\theta}^{(k)}$:

$$\mathbf{y}_{\text{rep}}^{(k)} \sim p(\mathbf{y}_{\text{rep}}|\boldsymbol{\theta}^{(k)}) \tag{3.3}$$

3c: Calculate the observed discrepancy $D(\mathbf{y}_{\text{obs}}; \boldsymbol{\theta}^{(k)})$ and the replicated discrepancy $D(\mathbf{y}_{\text{rep}}^{(k)}; \boldsymbol{\theta}^{(k)})$.

3d: Repeat Steps 3a to 3c for $k = 1, \ldots, K$ (e.g., $K = 1000$).

Step 4: Compute the ppp as the proportion of replicated data sets where $D(\mathbf{y}_{\text{rep}}^{(k)}; \boldsymbol{\theta}^{(k)})$ was greater than or equal to $D(\mathbf{y}_{\text{obs}}; \boldsymbol{\theta}^{(k)})$:

$$\text{ppp} = K^{-1} \sum_{k=1}^{K} I(D(\mathbf{y}_{\text{rep}}^{(k)}; \boldsymbol{\theta}^{(k)}) \geqslant D(\mathbf{y}_{\text{obs}}; \boldsymbol{\theta}^{(k)})), \tag{3.4}$$

where the indicator function $I(\cdot)$ equals 1 if the replicated discrepancy $D(\mathbf{y}_{\text{rep}}^{(k)}; \boldsymbol{\theta}^{(k)})$ is larger than or equal to the observed discrepancy $D(\mathbf{y}_{\text{obs}}; \boldsymbol{\theta}^{(k)})$, and 0 otherwise.

Hence the goal of the algorithm is to obtain a large set of K draws from the posterior (Step 3a), generate replicated data sets for all posterior draws (Step 3b), and compute the discrepancy based on the posterior draw for the observed data and the replicated data based on all posterior draws (Step 3c). This results in a set of K pairs of observed discrepancies and replicated discrepancies. The ppp is defined as the proportion of draws where the replicated discrepancy is larger than the observed discrepancy (Step 4). As an example Figure 3.1 displays $K = 1,000$ pairs of observed and replicated discrepancies in a posterior predictive check for independence of 4 dichotomous variables with a Pearson χ^2 discrepancy measure (elaborated in the next section). The ppp is equal to the proportion of pairs where replicated discrepancies are at least as large as the observed discrepancies (above the line where Drep = Dobs). In this example, the ppp was equal to .146.

Note that the prior that is specified in Step 1 does not play an important role when computing the ppp because typically the prior is completely dominated by the likelihood. It is even possible to specify a non-informative improper prior as long as the resulting posterior in Step 2 is proper.

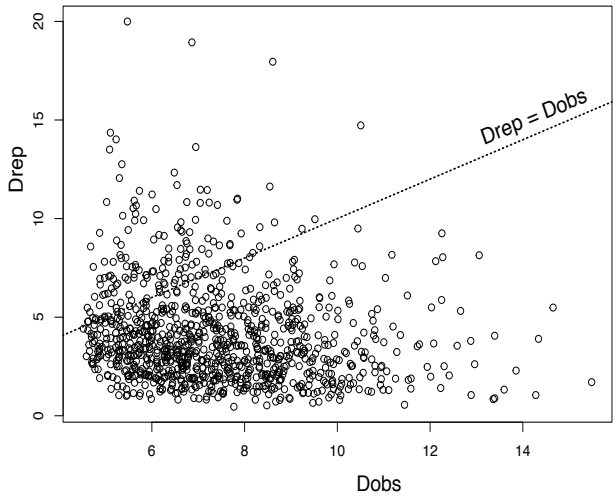

Figure 3.1: Plot of pairs of observed and replicated discrepancies, $D(\mathbf{y}_{\text{obs}}; \boldsymbol{\theta}^{(k)})$ (Dobs) and $D(\mathbf{y}_{\text{rep}}^{(k)}; \boldsymbol{\theta}^{(k)})$ (Drep), for $k = 1, \ldots, 1000$, when testing independence of 4 dichotomous variables using a Pearson χ^2 discrepancy measure. The pairs were obtained using Algorithm 1. The ppp is defined as the proportion of pairs where the replicated discrepancies are at least as large as the observed discrepancies, which was equal to .146.

3.2.1 Prior-calibrated posterior predictive p values.

It has been shown in previous work that the ppp is generally non-uniform under the null model (Hjort et al., 2006; Meng, 1994; Robins et al., 2000; van Kollenburg et al., 2015). A solution to the non-uniformity of the ppp has been proposed in which the ppp is calibrated with respect to a proper informative prior of the parameters (Hjort et al., 2006). This proper prior is used to construct a reference distribution for the ppp to check how extreme the observed ppp is. We shall refer to the resulting ppp as the prior-calibrated ppp (prior-cppp). The prior-cppp is uniformly distributed under the null model and the chosen proper prior, and therefore results in accurate type I error rates under the null. The prior-cppp can be obtained as follows.

Algorithm 3.2: Computation of the prior-cppp

Step 1: Specify an informative proper prior, $p(\boldsymbol{\theta})$, based on one's prior knowledge and choose a discrepancy measure, $D(\mathbf{y}; \boldsymbol{\theta})$.

Step 2: Compute the ppp for the observed data using Algorithm 1. The observed ppp will be denoted by ppp_{obs}.

Step 3: Obtain a reference distribution of the ppp using the informative prior in Step 1:

3a: Draw a parameter value, $\boldsymbol{\theta}^{(l)}_{\text{prior}}$, from the informative prior:

$$\boldsymbol{\theta}^{(l)}_{\text{prior}} \sim p(\boldsymbol{\theta}).$$

3b: Sample a data set using the likelihood of the model given the prior draw:

$$\mathbf{y}^{(l)}_{\text{prior}} \sim p(\mathbf{y}_{\text{prior}}|\boldsymbol{\theta}^{(l)}_{\text{prior}}) \tag{3.5}$$

3c: Compute the corresponding $\text{ppp}^{(l)}$ for data set $\mathbf{y}^{(l)}_{\text{prior}}$ using Algorithm 1.

3d: Repeat Steps 3a to 3c for $l = 1, \ldots, L$ (e.g., $L = 1000$).

Step 4: Calculate the prior-cppp as the proportion of $\text{ppp}^{(l)}$'s that are smaller than or equal than the observed ppp:

$$\text{prior-cppp} = L^{-1} \sum_{l=1}^{L} I(\text{ppp}_{\text{obs}} \geqslant \text{ppp}^{(l)}), \tag{3.6}$$

where the indicator function $I(\cdot)$ equals 1 if the observed ppp (ppp_{obs}) is larger than or equal to the prior-based ppp ($\text{ppp}^{(l)}$), and 0 otherwise.

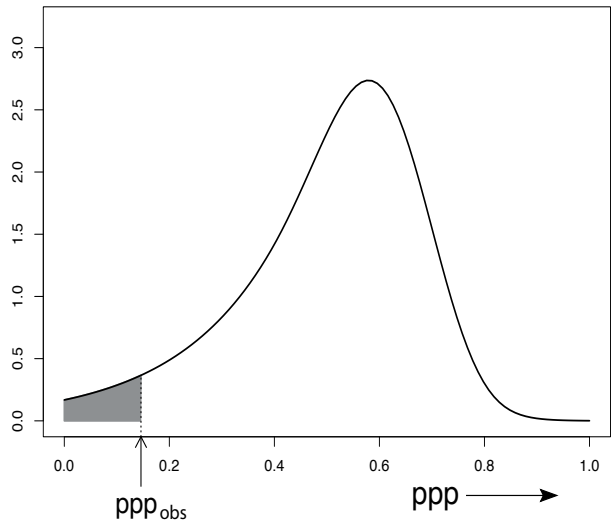

Figure 3.2: Reference distribution of prior-based ppp's, ppp$^{(l)}$, for a test of independence of 4 dichotomous variables using uniform priors for the response probabilities, for $l = 1, \ldots, 1000$ (Algorithm 2). The prior-cppp is estimated as the proportion of prior-based ppp's that are smaller than the observed ppp (grey area). The prior-cppp was equal to .034.

Hence in the posterior predictive check using the prior-cppp, the ppp itself is treated as a test statistic (Step 4) where the informative prior from Step 1 is used to obtain a reference distribution.

Figure 3.2 shows the reference distribution of prior-based ppp's, ppp$^{(l)}$ (obtained in Step 3d of Algorithm 2), for the same test and data that resulted in the ppp in Figure 3.1. Uniform priors were used for the response probabilities of the 4 variables. The ppp of the observed data was equal to ppp$_{\text{obs}} = .146$. The prior-cppp is estimated as the proportion of prior-based ppp's that are smaller than the ppp of the observed data (Step 4 in Algorithm 2; grey area in Figure 3.2). In this example the prior-cppp was equal to .034.

A central property of the posterior predictive check based on the prior-cppp is that it simultaneously tests model fit and prior fit for the observed data. This implies that the prior-cppp may result in a rejection of the null

model either in the case of model misfit (i.e., the model does not fit the observed data) or in the case of prior misfit (i.e., one's prior beliefs about the unknown parameters conflict with the information in the observed data). As a consequence, the prior-cppp may highly depend on the chosen prior. This will be shown in the next section. We argue that the prior-cppp should only be used if a researcher has clear prior information that can be translated to an informative prior for all model parameters, and if the researcher is also interested in testing these prior beliefs simultaneously with the statistical model.

3.2.2 Posterior-calibrated posterior predictive p values.

In the absence of prior information or when one is only interested in evaluating model misfit with accurate type I error rates, neither the prior-cppp nor the standard ppp can be used. To keep the useful properties of the ppp (i.e., the flexibility to detect any form of model misfit, straightforward computation using standard MCMC algorithms, and its usage with non-informative (improper) priors), while maintaining an accurate type I error rate if the null model is true, we propose to calibrate the ppp under the posterior under the null model. The resulting ppp will be referred to as the posterior-calibrated ppp (posterior-cppp). The exact steps to compute the posterior-cppp are given in Algorithm 3.

Algorithm 3.3: Computation of the posterior-cppp

Step 1: Specify a prior, $p(\boldsymbol{\theta})$, and choose a discrepancy measure, $D(\mathbf{y}; \boldsymbol{\theta})$.

Step 2: Derive the posterior via (3.1) and compute the ppp for the ob-

served data using Algorithm 1. The observed ppp will be denoted by ppp_{obs}.

Step 3: Obtain a reference distribution of the ppp using the posterior from Step 2:

3a: Draw a parameter value, $\boldsymbol{\theta}^{(m)}_{\text{post}}$, from the posterior:

$$\boldsymbol{\theta}^{(m)}_{\text{post}} \sim p(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}).$$

3b: Sample a data set using the likelihood of the model given the posterior draw:

$$\mathbf{y}^{(m)}_{\text{post}} \sim p(\mathbf{y}_{\text{post}}|\boldsymbol{\theta}^{(m)}_{\text{post}}) \tag{3.7}$$

3c: Compute the corresponding $\text{ppp}^{(m)}$ for data set $\mathbf{y}^{(m)}_{\text{post}}$ using Algorithm 1.

3d: Repeat Steps 3a to 3c for $m = 1, \ldots, M$ (e.g., $M = 1000$).

Step 4: Calculate the posterior-cppp as the proportion of $\text{ppp}^{(m)}$'s that are smaller than or equal to the observed ppp:

$$\text{posterior-cppp} = M^{-1} \sum_{m=1}^{M} I(\text{ppp}_{\text{obs}} \geqslant \text{ppp}^{(m)}), \tag{3.8}$$

where the indicator function $I(\cdot)$ equals 1 if the constraint is satisfied, and 0 otherwise.

Note that the only difference between Algorithm 2 for the prior-cppp and Algorithm 3 for the posterior-cppp is in the generation of parameters draws for $\boldsymbol{\theta}$ where either an informative prior is used (Step 3a in Algorithm 2) or the posterior is used (Step 3a in Algorithm 3). When computing

the posterior-cppp, we recommend to use diffuse or non-informative priors that are completely dominated by the data. Note that if an informative prior would be used for the computation of the posterior-cppp, the testing criterion would become a hybrid of the prior-cppp and the posterior-cppp. Though one could imagine very specific situations in which this may be useful, researchers typically compute ppp's using diffuse priors (Choi et al., 2010; Berkhof et al., 2003; Hoverd & Sibley, 2013).

Figure 3.3 shows the reference distribution of posterior-based ppp's, $\text{ppp}^{(m)}$ (obtained in Step 3c of Algorithm 3), for the same test and data that resulted in the ppp and the prior-cppp in Figures 3.1 and 3.2, respectively. Uniform priors were used for the response probabilities of the 4 variables. The posterior-cppp is estimated as the proportion of posterior-based ppp's that are smaller than the ppp of the observed data (Step 4 in Algorithm 3; grey area in Figure 3.3). In this example the posterior-cppp was equal to .018.

Next we will investigate the performance of the different posterior predictive checks in various testing problems by looking at frequentist criteria such as type I error rates and power.

3.3 Application I: a Bayesian Test for Independence

The first application is a simple test of independence of J dichotomous variables (outcome 1 or 2). The goal of the application is (i) to illustrate how a posterior predictive check can be conducted using the three different types of ppp's, (ii) to get insight about the type I error rates of the different posterior predictive checks for this simple test, and (iii) to investigate to

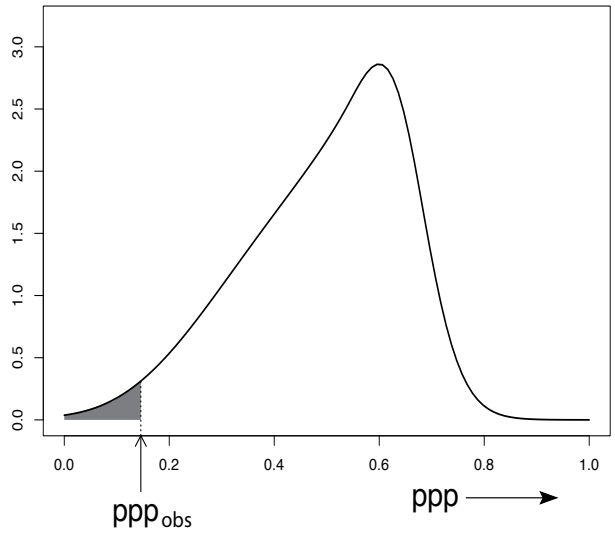

Figure 3.3: Reference distribution of posterior-based ppp's, $\text{ppp}^{(m)}$, for a test of independence of 4 dichotomous variables using uniform priors for the response probabilities, for $m = 1, \ldots, 1000$ (Algorithm 3). The posterior-cppp is estimated as the proportion of posterior-based ppp's that are smaller than the observed ppp (grey area). The posterior-cppp was equal to .018.

what degree the prior-cppp depends by the choice of the proper prior. This will be assessed by means of a Monte Carlo study.

The data consists of responses of N individuals to J dichotomous variables. We are interested in the following test:

$\mathcal{M}_0$: The J dichotomous variables are independent.

$\mathcal{M}_1$: Not $\mathcal{M}_0$, i.e., there is a dependence between at least two variables.

It is standard practice to specify conjugate priors with independent beta distributions for the response probabilities, denoted by π_{1j}, for variables $j = 1, \ldots, J$. The beta prior will be written as $Beta(\pi_{1j}|\alpha_j, \beta_j)$, where α_j and β_j are the hyper parameters discussed in the following subsection. In the case of independent variables, as under $\mathcal{M}_0$, the likelihood follows

a binomial distribution, resulting in a posterior with independent beta distributions, $Beta(\pi_{1j}|n_{1j}+\alpha_j, N-n_{1j}+\beta_j)$, where n_{1j} is the number of individuals having response 1 to variable j (for technical details see Appendix 3.A).

We cross-tabulate the variables resulting in a contingency table with $S=2^J$ cells. Thus cell s corresponds to a particular response pattern $\mathbf{y}_s$, for $s=1,\ldots,S$, e.g., $\mathbf{y}_1$ is a vector of J ones. Under $\mathcal{M}_0$ the probability of pattern $\mathbf{y}_s$, denoted by π_s (with a slight abuse of notation), is given by

$$\pi_s = \prod_{j=1}^{J} (\pi_{1j})^{d_{js}} (1-\pi_{1j})^{1-d_{js}}, \tag{3.9}$$

where the dummy indicator d_{js} equals 1 if the response to variable j in pattern s is 1, and 0 otherwise. For example the response probability of pattern $\mathbf{y}_1$ with J ones equals $\pi_1 = \pi_{11} \times \ldots \times \pi_{1J}$. The number of individuals in the data having response pattern s will be denoted by n_s.

To assess overall model fit of $\mathcal{M}_0$ we can use a discrepancy measure based on the Pearson χ^2-statistic, given by

$$D_{\chi^2}(\mathbf{n};\boldsymbol{\pi}) \;=\; \sum_{s=1}^{S} \frac{n_s - e_s}{e_s}, \tag{3.10}$$

where e_s is the expected number of individuals having response probability s based on the pattern probabilities in $\boldsymbol{\pi}$, i.e., $e_s = N\pi_s$.

To obtain the k-th posterior draw for the pattern probabilities $\boldsymbol{\pi}$, first draw the $\pi_{1j}^{(k)}$'s from their beta posteriors, and subsequently, plug these draws in (3.9) to obtain $\boldsymbol{\pi}^{(k)}$ (Step 3a in Algorithm 1). Given the k-th posterior draw, the k-th replicated data set can be drawn from the Multinomial$(y_{\text{rep},1},\ldots,y_{\text{rep},S}|\pi_1^{(k)},\ldots,\pi_S^{(k)})$ distribution (Step 3b in Algo-

rithm 1). Subsequently, the observed and replicated discrepancies (Step 3c of Algorithm 1) are calculated as

$$D_{\chi^2}(\mathbf{n}_{\text{obs}}; \boldsymbol{\pi}^{(k)}) = \sum_{s=1}^{S} \frac{(n_{\text{obs},s} - e_s^{(k)})^2}{e_s^{(k)}} \tag{3.11}$$

$$D_{\chi^2}(\mathbf{n}_{\text{rep}}^{(k)}; \boldsymbol{\pi}^{(k)}) = \sum_{s=1}^{S} \frac{(n_{\text{rep},s}^{(k)} - e_s^{(k)})^2}{e_s^{(k)}}, \tag{3.12}$$

in which $n_{\text{obs},s}$ and $n_{\text{rep},s}^{(k)}$ are the frequencies of pattern s in the observed data and the k-th replicated data, respectively.

3.3.1 Simulation set-up.

The type I error rates of the ppp, prior-cppp and posterior-cppp were investigated under conditions with $J = 4$ independent dichotomous variables and where the success probabilities π_{1j} in population A are equal to .2 and in population B follow a $Beta(6, 24)$-distribution, for all variables $j = 1, \ldots, J$. Sample sizes of $N = 100$ and $N = 1000$ were considered. Under each condition, 2000 data sets were generated.

For the ppp, non-informative uniform priors were used for the response probabilities π_{1j} by setting the hyper parameters to $\alpha_j = \beta_j = 1$, for all $j = 1 \ldots, J$ (Step 1 of Algorithm 1). To compute the ppp, the number of replicated data sets was set to $K = 1,000$ (Step 3 of Algorithm 1). To compute the posterior-cppp also non-informative uniform priors were used for the response probabilities (Step 1 of Algorithm 3).

Three different prior-cppp's were considered based on three different priors (Step 1 of Algorithm 2).

1. Prior 1: $\pi_{1j} \sim Beta(6, 24)$ for all $J = 4$ variables (Figure 3.4, dotted line). This prior is *in agreement* with population B, and therefore

results in prior-cppp's with a uniform distribution in this case. Because this prior is concentrated around .2, it is expected that the sampling distribution of the prior-cppp's is also close to uniform for data generated from population A where $\pi_{1j} = .2$.

2. Prior 2: $\pi_{1j} \sim Beta(15, 15)$ for all $J = 4$ variables (Figure 3.4, dashed line). This prior is concentrated around .5 and therefore this prior is *not* in agreement with both populations. It is expected that the sampling distributions of this prior-cppp's are not uniformly distributed under population A and B.

3. Prior 3: $\pi_{1j} \sim Beta(1, 1)$ for all $J = 4$ variables (Figure 3.4, solid line). This uniform prior assumes that every probability value is equally likely. This is a standard prior choice if no prior information is available.

As noted earlier, the prior-cppp simultaneously tests the employed statistical model and the informative prior. Thus when using the prior-cppp the test can be formulated as

$\mathcal{M}_0$: The J dichotomous variables are independent, and the response probabilities for variable j follow a $\pi_{1j} \sim Beta(\alpha_j, \beta_j)$-prior, for $j = 1, \ldots, J$.

$\mathcal{M}_1$: Not $\mathcal{M}_0$, i.e., there is a dependency between at least two variables and/or the priors under $\mathcal{M}_0$ are not in accordance with the information in the data.

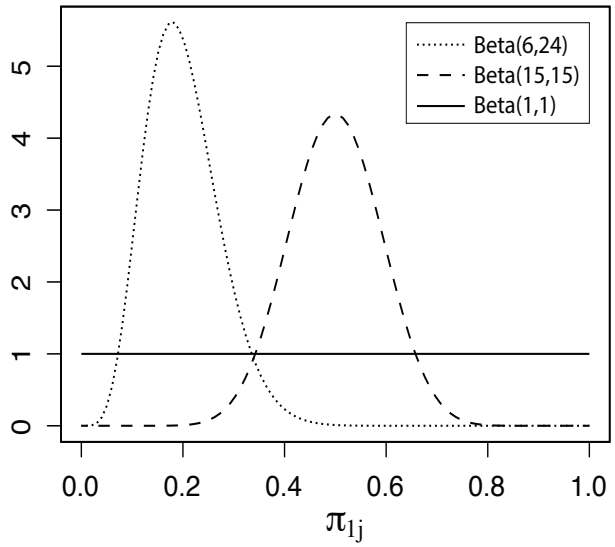

Figure 3.4: Three beta-priors for the probability of a success to variable j.

3.3.2 Results of the Monte Carlo study.

The type I error rates, which were based on the common significance level of $\alpha = .05$, can be found in the first row of Table 3.1. The corresponding Monte Carlo errors were computed as $\sqrt{\frac{\hat{p}(1-\hat{p})}{2000}}$, where $\hat{p}$ corresponds to the estimated type I error rates and the denominator corresponds to 2000 because the estimate is based on 2000 randomly generated data sets. As can be seen, the type I error rates for the ppp are around .002, which is much too low. This can be understood by looking at the sampling distribution of the ppp's which are plotted in Figure 3.5 for each of the four scenarios (dotted lines). As can be seen the sampling distributions of the ppp's are peaked around .55, which explains the dramatically low type I error rates. As a consequence the ppp test is too conservative.

Table 3.1 (second row, last two columns) shows that the type I error rates of the prior-cppp are accurate when the prior is correctly specified, i.e., when using $Beta(6, 24)$-priors in the case of $Beta(6, 24)$-distributions in the populations. Also the corresponding sampling distributions are approximately uniform (Figure 3.5; right panels, dash-dotted lines). In all

Table 3.1: Type I error rates with Monte Carlo errors for the ppp, three prior-cppp's based on a $Beta(6, 24)$-prior, a $Beta(15, 15)$-prior, and a $Beta(1, 1)$-prior, and the posterior-cppp.

Population	$\pi_{1j} = .2$		$\pi_{1j} \sim Beta(6, 24)$	
Sample size	$N = 100$	$N = 1000$	$N = 100$	$N = 1000$
ppp	$.002 \pm .001$	$.002 \pm .001$	$.002 \pm .001$	$.000 \pm .000$
prior-cppp with $Beta(6, 24)$-prior	$.030 \pm .004$	$.043 \pm .005$	$.042 \pm .004$	$.059 \pm .005$
prior-cppp with $Beta(15, 15)$-prior	$.643 \pm .011$	$.043 \pm .005$	$.646 \pm .011$	$.037 \pm .004$
prior-cppp with $Beta(1, 1)$-prior	$.023 \pm .003$	$.030 \pm .004$	$.029 \pm .004$	$.030 \pm .004$
posterior-cppp	$.043 \pm .005$	$.048 \pm .005$	$.047 \pm .005$	$.047 \pm .005$

other situations, the prior-cppp neither results in accurate type I error rates nor in approximately uniform sampling distributions. This is even the case when calibrating the prior-cppp under standard uniform priors for the response probabilities (Table 3.1; $Beta(1, 1)$ in the fourth row). Furthermore, it is interesting to observe that the rejection rates for the prior-cppp are lower than 5% in most case when the prior does not correspond with the population distribution (except when using the $Beta(15, 15)$-prior and $N = 100$). This is somewhat surprising because one might expect a larger rejection rate than 5% due to the mismatch of the prior.

The results for the posterior-cppp are all acceptable. As can be seen in Figure 3.5 (thick solid lines), the sampling distributions are approximately uniform in all scenarios. Furthermore, the last row in Table 3.1 shows that the type I error rates of the posterior-cppp do not differ significantly from .05 in all scenarios.

In sum, this simple example provides the following useful insights about posterior predictive checking. First, the standard ppp, even though it is flexible and simple to compute, does not result in accurate type I error rates, not even in this very simple test for independence. Second, the prior-cppp highly depends on the exact choice of the specified proper prior. For this reason we cannot recommend the prior-cppp for default model

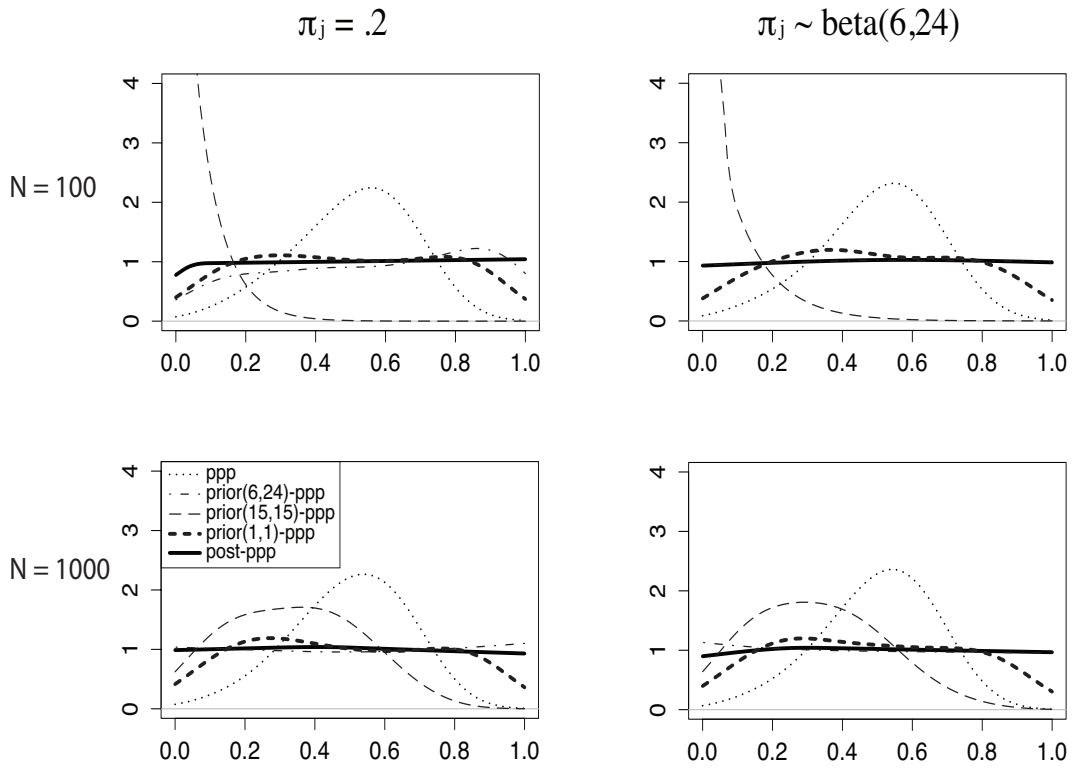

Figure 3.5: The sampling distribution of different posterior predictive p values (ppp's) in the case of $J = 4$ dichotomous variables, $N = 100$ observations (upper panels) and $N = 1000$ (lower panels), and a true population where $\pi_{1j} = .2$ and $\pi_{1j} \sim Beta(6, 24)$, $j = 1, \ldots, 4$. Distributions are displayed for the standard ppp using a uniform prior for π_{1j} (dotted line), three different prior-cppp's based on a $Beta(6, 24)$-prior (dash-dotted line), a $Beta(15, 15)$-prior (thin dashed line), and a $Beta(1, 1)$-prior (thick dashed line), and the posterior-cppp using a uniform prior (thick solid line).

checking. Therefore, the prior-cppp will not be considered further in this chapter. Third, the posterior-cppp with standard diffuse priors clearly outperforms the ppp and the prior-cppp by providing accurate type I error rates.

3.4 Application II: Testing for Extreme Observations in Regression

To illustrate the generality of the proposed approach, we evaluate the type I error rates and power of the posterior-cppp in a regression model. The standard linear regression model assumes that a dependent variable can be explained by a linear combination of certain predictor variables and a normally distributed error. The model can be written as

$$y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2), \tag{3.13}$$

for $i = 1, \ldots, N$, where y_i is the i-th observation of the dependent variable, $\mathbf{x}_i$ is a vector with k predictor variables of the i-th observation, $\boldsymbol{\beta}$ is a vector with k unknown regression coefficients, and σ^2 is the error variance. The standard independence Jeffreys prior is used for the unknown parameters $(\boldsymbol{\beta}, \sigma^2)$ (Step 1 of Algorithm 1 and 3). The posterior follows a normal-inverse-gamma distribution (Step 2 of Algorithm 1; Step 3 of Algorithm 3). Technical details can be found in Appendix 3.B

To illustrate the flexibility of posterior predictive checking we shall test whether the employed linear regression model appropriately captures extreme observations. This can be achieved using the following discrepancy measure (Hjort et al., 2006):

$$D_{\max}(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) = \max_{i \in \{1,\ldots,N\}} |y_i - \mathbf{x}_i'\boldsymbol{\beta}|/\sigma, \tag{3.14}$$

where $\mathbf{y} = (y_1, \ldots, y_N)'$ is the vector containing the N observations of the dependent variable, and $\mathbf{X}$ is a $N \times k$ matrix where the i-th row contains the k predictor variables, denoted by $\mathbf{x}_i$. The discrepancy measure com-

putes the largest standardised error between the observations, y_i, and their predictions according to the model, i.e., $\mathbf{x}_i'\boldsymbol{\beta}$. Note that we are not interested in determining which observations are extreme (i.e., we are not doing outlier detection); we are only interested in checking whether the employed linear regression model appropriately captures extreme observations.

For a given posterior draw $(\boldsymbol{\beta}^{(k)}, \sigma^{2,(k)})$, a replicated data set $\mathbf{y}_{\text{rep}}^{(k)}$ can be obtained via (3.13) using the matrix of covariates from the observed data, $\mathbf{X}_{\text{obs}}$. Note that is also standard practice to assume the covariates to be fixed in Bayesian linear regression. The observed and replicated discrepancies are then given by

$$D_{\max}(\mathbf{y}_{\text{obs}}, \mathbf{X}_{\text{obs}}; \boldsymbol{\beta}^{(k)}, \sigma^{2,(k)}) = \max_{i \in \{1,\ldots,N\}} |y_{\text{obs},i} - \mathbf{x}_{\text{obs},i}'\boldsymbol{\beta}^{(k)}|/\sigma^{(k)}, \tag{3.15}$$

$$D_{\max}(\mathbf{y}_{\text{rep}}^{(k)}, \mathbf{X}_{\text{obs}}; \boldsymbol{\beta}^{(k)}, \sigma^{2,(k)}) = \max_{i \in \{1,\ldots,N\}} |y_{\text{rep},i}^{(k)} - \mathbf{x}_{\text{obs},i}'\boldsymbol{\beta}^{(k)}|/\sigma^{(k)}. \tag{3.16}$$

3.4.1 Monte Carlo study when testing extreme observations.

A Monte Carlo simulation was conducted to investigate whether the ppp and the posterior-cppp are able to pick up extreme observations using the discrepancy measure in Equations (3.15) and (3.16). Two different forms of misfit were considered. First, instead of a normal distribution, errors were generated using a Student t distribution with 1, 2, 5, 20, and 50 degrees of freedom. Note that when the degrees of freedom equals ∞ the errors are normally distributed, while 1 degree of freedom corresponds to a Cauchy distribution resulting in much more extreme observations than normally distributed errors. Second, the residual standard deviation was set to be a monotonic function of the sum of the explanatory variables,

which results in heteroskedastic errors. The residual standard deviation for the i-th observation was set to $\sigma_i = 1 + c \times \frac{w_i - w_{\min}}{w_{\max} - w_{\min}}$, where $w_i = \mathbf{x}_i'\mathbf{1}$ is the sum of the explanatory variables, and $w_{\min} = \min_i\ w_i$ and $w_{\max} = \max_i\ w_i$. Larger values for c imply larger error variances for large values of the explanatory variables. Note that $c = 0$ implies homoskedastic errors of $N(0, 1)$. For the simulation we choose the following values $c =$ 1, 2, 3, 5, and 10. Sample sizes were set to $n =$ 50, 100, and 250. For every condition, 1000 datasets were generated using 3 explanatory variables from independent standard normal distributions and regression coefficients equal to $\boldsymbol{\beta} = (.3, .3, .3)$.

The results on the type I error rates and the power can be found in Table 3.2. Again these results show that the ppp is too conservative with error rates that are too small for all sample sizes. The posterior-cppp on the other hand results in reasonable type I error rates. Furthermore, the posterior-cppp consistently has higher power than the ppp in the case of model misfit.

3.4.2 An empirical analysis of quality-of-life in elderly.

A posterior predictive check was performed to detect model misfit for a regression model testing the effects of physical, psychological and social frailty on the quality of life with respect to social relationships of elderly people (Age, mean±SD 84.8±9.7, range 55–101)(Gobbens, Krans, & van Assen, 2015). The sample consisted of $n = 156$ observations. The regression model consisted of the three explanatory variables of frailty, presence of disease, and 8 control variables (such as marital status and income), as well as an intercept.

Table 3.2: Estimated type I error rates (first row), power in the case of Student $t(\nu)$ distributed errors with degrees of freedom ν (second to sixth row), and power in the case of heterogeneous normally distributed errors with $\sigma_i = 1 + c \times \frac{w_i - w_{\min}}{w_{\max} - w_{\min}}$.

	Sample Size					
	50		100		250	
$\epsilon \sim$	ppp	post-cppp	ppp	post-cppp	ppp	post-cppp
$N(0,1)$	$.004 \pm .002$	$.045 \pm .007$	$.015 \pm .004$	$.041 \pm .006$	$.034 \pm .006$	$.058 \pm .007$
t(50)	$.007 \pm .003$	$.073 \pm .008$	$.028 \pm .005$	$.073 \pm .008$	$.053 \pm .007$	$.079 \pm .009$
t(20)	$.016 \pm .004$	$.078 \pm .008$	$.059 \pm .007$	$.141 \pm .011$	$.127 \pm .011$	$.161 \pm .012$
t(5)	$.168 \pm .012$	$.345 \pm .015$	$.388 \pm .015$	$.553 \pm .016$	$.710 \pm .014$	$.779 \pm .013$
t(2)	$.609 \pm .015$	$.798 \pm .013$	$.906 \pm .009$	$.947 \pm .007$	$.998 \pm .001$	$.998 \pm .001$
t(1)	$.928 \pm .008$	$.977 \pm .005$	$.998 \pm .001$	$.998 \pm .001$	$1.000 \pm .000$	$1.000 \pm .000$
$c = 1$	$.045 \pm .007$	$.184 \pm .012$	$.103 \pm .010$	$.226 \pm .013$	$.204 \pm .013$	$.283 \pm .014$
$c = 2$	$.086 \pm .009$	$.316 \pm .015$	$.244 \pm .014$	$.390 \pm .015$	$.413 \pm .016$	$.488 \pm .016$
$c = 3$	$.144 \pm .011$	$.408 \pm .016$	$.319 \pm .015$	$.480 \pm .016$	$.548 \pm .016$	$.621 \pm .015$
$c = 5$	$.196 \pm .013$	$.459 \pm .016$	$.422 \pm .016$	$.617 \pm .015$	$.640 \pm .015$	$.727 \pm .014$
$c = 10$	$.270 \pm .014$	$.569 \pm .016$	$.528 \pm .016$	$.687 \pm .015$	$.778 \pm .013$	$.852 \pm .011$

The ppp and the posterior-cppp were computed with the discrepancy measure in Equation (3.14) using the standard independence Jeffreys prior. The ppp calculated with $K = 500$ replications was equal to .064, which would generally not be considered to indicate significant misfit. The posterior-cppp with $M = 500$, on the other hand, was equal to .012, which would in most situations be considered to indicate significant misfit. For this reason, it is recommendable to reconsider the employed regression model before making inferences about the quality of life of elderly people.

3.5 Application III: Bayesian Tests for Latent Class Analysis

To get more insights about the performance of posterior predictive checks in more complex situations, we shall test model misfit of latent class models. Latent class models are commonly used to create typologies or clusterings

of observations, based on their response patterns (Goodman, 1974). It has been shown that p values based on asymptotic sampling distributions, p values based on the parametric bootstrap, and posterior predictive p values based on test statistics may not result in accurate type I error rates for this type of model (van Kollenburg et al., 2015). For this reason, the latent class model is a good test case to check the performance of the posterior-cppp.

Again note that in order to apply the prior-cppp, informative priors need to be formulated for all the unknown model parameters in the latent class model, such as the latent class proportions and the response probabilities for a given latent class. This is not feasible from a practical point of view. Based on our experience researchers are mainly interested in evaluating the fit of a latent class model, and not in evaluating the prior beliefs about the unknown model parameters. These considerations again exemplify the limited usability of the prior-cppp in more complex models. Therefore the prior-cppp will not be considered in this application.

We shall consider a latent class model for a given data set of N individuals who responded to J dichotomous variables, when assuming the individuals can be divided into C latent classes. Appendix 3.C contains technical details of the latent class model. Two Monte Carlo studies will be conducted: (i) testing the assumption of local independence, and (ii) testing for the number of latent classes. Additionally, the posterior-cppp will be used for an empirical latent class analysis on sub-types of depression in males.

3.5.1 Monte Carlo study on bivariate residuals

Let us first focus on the key assumption of the latent class model that the observations of each pair of variables are independent given the (unknown)

latent class memberships of the individuals. To test whether this property is violated for variable pair (j, j'), we use the bivariate residual (BVR) (Vermunt & Magidson, 2016):

$$D_{BVR_{jj'}}(\mathbf{n}; \boldsymbol{\rho}, \boldsymbol{\pi}) = \sum_{s=1}^{4} \frac{(n_s - e_s)^2}{e_s}, \tag{3.17}$$

where the sum is over the $S = 2^2 = 4$ cells of the contingency table, n_s is the observed frequency of response pattern s, and $e_s = N\pi_s$ denotes the expected frequency of response pattern s of the variable pair (j, j') given the latent class proportions $\boldsymbol{\rho}$ and the response probabilities $\boldsymbol{\pi}$ of the latent class model (see Appendix 3.C).

For a given posterior draw of the latent class proportions, $\boldsymbol{\rho}^{(k)}$, and response probabilities, $\boldsymbol{\pi}^{(k)}$, the observed and replicated discrepancies are calculated as

$$D_{BVR_{jj'}}(\mathbf{n}_{\text{obs}}; \boldsymbol{\rho}^{(k)}, \boldsymbol{\pi}^{(k)}) = \sum_{s=1}^{4} \frac{(n_{\text{obs},s} - e_s^{(k)})^2}{e_s^{(k)}}, \tag{3.18}$$

$$D_{BVR_{jj'}}(\mathbf{n}_{\text{rep}}^{(k)}; \boldsymbol{\rho}^{(k)}, \boldsymbol{\pi}^{(k)}) = \sum_{s=1}^{4} \frac{(n_{\text{rep},s}^{(k)} - e_s^{(k)})^2}{e_s^{(k)}}, \tag{3.19}$$

where $\mathbf{n}_{\text{rep}}^{(k)}$ denotes the frequencies of the response patterns of a replicated data set generated using $(\boldsymbol{\rho}^{(k)}, \boldsymbol{\pi}^{(k)})$ (see Appendix 3.C).

A Monte Carlo simulation was conducted to evaluate the power and type I error rates of the posterior-cppp when testing conditional independence for pairs of variables using the BVR. We generated $J = 6$ dichotomous variables from a latent class model with $C = 2$ equally sized classes (i.e., $\rho_1 = \rho_2 = .5$). Sample sizes were either $N = 100$, 500 or 1000. The conditional probabilities for a 1-response were $\pi_{1jc} = .8$ in class 1 and

$\pi_{1jc} = .2$ in class 2 for variables $j = 1$ to 5. For class 1, the probability of 1-response to variable 6 conditional on having a 1-response to variable 5 was set to $(\pi_{161}|y_5 = 1) = .8 + \delta$, where δ was used to add conditional association/local dependence between the last two variables. For observations with a 2-response on variable 5, we set $(\pi_{161}|y_5 = 2) = .8$. In class 2, all response probabilities were set to the complement of the response probability of class 1, i.e., $\pi_{rj2} = 1 - \pi_{rj1}$. Thus the 2-class model fits when $\delta = 0$, while local independence is violated between variables 5 and 6 in the 2-class model when $\delta \neq 0$. Table 3.3 summarises the response probabilities for all variables.

Under each condition 500 datasets were generated. The original ppp and the posterior-cppp were computed using vague uniform $Beta(1, 1)$-priors for all model parameters (i.e., class proportion ρ_c, and the probabilities of a 1-response in each class, π_{1jc}, for $j = 1, \ldots, 6$, and $c = 1$, or 2). To compute the original ppp, $K = 500$ replications were generated. To compute the posterior-cppp, a reference distribution was obtained using $M = 500$ posterior-based data sets. To compute each ppp for the reference distribution $K = 501$ replications were used, so that there are no ties with the observed ppp.

The results for the study on bivariate residuals are displayed in Table 3.4. The condition in which $\delta = 0$ confirms that the posterior-cppp has accurate type I errors which is not the case for the standard ppp. Moreover, the results show that the power to detect misfit (in the current example being local dependency between pairs of variables) is greatly improved by calibrating the original ppp with respect to the posterior distribution, as is done in the posterior-cppp.

Table 3.3: Class proportions ρ_c and response probabilities π_{rjc} for response r on dichotomous variable j under class c, for $r = 1$ or 2, $j = 1, \ldots, 6$, and class $c = 1$ or 2. Note that $\pi_{2jc} = 1 - \pi_{1jc}$

class	$c = 1$	$c = 2$
ρ_c	.5	.5
π_{11c}	.8	.2
π_{12c}	.8	.2
π_{13c}	.8	.2
π_{14c}	.8	.2
π_{15c}	.8	.2
$\pi_{16c}\|y_5 = 1$	$.8 + \delta$	$.2 - \delta$
$\pi_{16c}\|y_5 = 2$	.8	.2

Table 3.4: Estimated type-I error rates (row where $\delta = 0$) and power (rows where $\delta \neq 0$) when testing local independence in a 2-class model using a significance level of .05.

	Sample size					
	100		500		1000	
δ	ppp	post-cppp	ppp	post-cppp	ppp	post-cppp
$-.2$	$.012 \pm .005$	$.186 \pm .017$	$.246 \pm .019$	$.912 \pm .013$	$.764 \pm .019$	$1.000 \pm .000$
$-.1$	$.000 \pm .000$	$.032 \pm .008$	$.000 \pm .000$	$.322 \pm .021$	$.016 \pm .006$	$.646 \pm .021$
$-.05$	$.000 \pm .000$	$.042 \pm .009$	$.000 \pm .000$	$.064 \pm .011$	$.000 \pm .000$	$.180 \pm .017$
.00	$.000 \pm .000$	$.058 \pm .010$	$.000 \pm .000$	$.048 \pm .010$	$.000 \pm .000$	$.040 \pm .009$
.05	$.000 \pm .000$	$.108 \pm .014$	$.000 \pm .000$	$.232 \pm .019$	$.000 \pm .000$	$.362 \pm .021$
.1	$.000 \pm .000$	$.208 \pm .018$	$.000 \pm .000$	$.548 \pm .022$	$.000 \pm .000$	$.840 \pm .016$
.2	$.000 \pm .000$	$.426 \pm .022$	$.000 \pm .000$	$.768 \pm .019$	$.000 \pm .000$	$.820 \pm .017$

3.5.2 Monte Carlo study on the number of latent classes.

The second major testing problems involves checking whether enough latent classes are specified. To do this we can assess the overall misfit of the latent class model. This can be done using the Pearson χ^2 or the likelihood

ratio G^2 as discrepancy measure, which are given by

$$D_{\chi^2}(\mathbf{n}; \boldsymbol{\rho}, \boldsymbol{\pi}) = \sum_{s=1}^{S} \frac{(n_s - e_s)^2}{e_s}, \tag{3.20}$$

and

$$D_{G^2}(\mathbf{n}; \boldsymbol{\rho}, \boldsymbol{\pi}) = 2 \sum_{s=1}^{S} n_s \ln(n_s/e_s), \tag{3.21}$$

where the sum is over all S possible response patterns, n_s denotes the observed frequency of response pattern s, and $e_s = N\pi_s$ denotes the expected frequency of response pattern s given the model parameters (Appendix 3.C).

Type I error rates of the ppp and the posterior-cppp for the discrepancy measures in (3.20) and (3.21) were investigated under a null model with two latent classes, by means of a Monte Carlo simulation. The population under the null model had $C = 2$ equally sized classes, with conditional response probabilities for a 1-response of $\pi_{1jc} = .8$ in class 1 and $\pi_{1jc} = .2$ in class 2 for all $J = 6$ dichotomous variables. The power of the tests was investigated by assuming a population with $C = 3$ equally sized classes with the same conditional probabilities for classes 1 and 2. For class 3, the probability of 1-response was $\pi_{1jc} = .8$ for the first half of the variables and $\pi_{1jc} = .2$ for the last half of the variables. Table 3.5 shows the parameter values for the conditions in which data was generated from a three-class model with six variables. Sample sizes were either $N = 100$, or 500.

For these conditions we ran 500 Monte Carlo simulations per condition. The ppp was calculated with $K = 500$ replications, and calibrated using $M = 500$ posterior-based data sets on which we performed a posterior predictive check with $K = 501$ replications. The ppp and the posterior-cppp were computed using default uniform $Beta(1, 1)$-priors.

Table 3.5: Class proportions ρ_c and response probabilities π_{rjc} for response 1 on dichotomous variable j under class c, for $r = 1$ or 2, $j = 1, \ldots, 6$, and class $c = 1$ or 2. Note that $\pi_{2jc} = 1 - \pi_{1jc}$

class	$c = 1$	$c = 2$	$c = 3$
ρ_c	1/3	1/3	1/3
π_{11c}	.8	.2	.8
π_{12c}	.8	.2	.8
π_{13c}	.8	.2	.8
π_{14c}	.8	.2	.2
π_{15c}	.8	.2	.2
π_{16c}	.8	.2	.2

Table 3.6: Estimated type-I error rates (rows where $C = 2$) and power (rows where $C = 3$) when testing the global fit of a two-class model when using a significance level of .05.

			χ^2		G^2	
C	N	π_{1j1}	ppp	$\text{cppp}_{\text{post}}$	ppp	$\text{cppp}_{\text{post}}$
2	100	.8	$.004 \pm .003$	$.054 \pm .010$	$.036 \pm .008$	$.062 \pm .011$
		.9	$.008 \pm .004$	$.028 \pm .007$	$.018 \pm .006$	$.052 \pm .010$
	500	.8	$.000 \pm .000$	$.038 \pm .009$	$.032 \pm .008$	$.050 \pm .010$
		.9	$.002 \pm .002$	$.046 \pm .009$	$.032 \pm .008$	$.046 \pm .009$
3	100	.8	$.402 \pm .022$	$.632 \pm .022$	$.372 \pm .022$	$.478 \pm .022$
		.9	$.416 \pm .022$	$.660 \pm .021$	$.372 \pm .022$	$.470 \pm .022$
	500	.8	$1.000 \pm .000$	$1.000 \pm .000$	$1.000 \pm .000$	$1.000 \pm .000$
		.9	$1.000 \pm .000$	$1.000 \pm .000$	$1.000 \pm .000$	$1.000 \pm .000$

The results for the study on the number of classes can be found in Table 3.6. The conditions in which the fitted model holds (when the true number of classes equals 2) confirm that the posterior-cppp has accurate type I error rates, unlike the ppp. Moreover, the posterior-cppp clearly has more power than the ppp to detect model misfit.

3.5.3 An empirical analysis of sub-types of depression in males.

The posterior-cppp was used to analyse the depression scale data for white male respondents from the problems of everyday life study (Schaeffer, 1988; Pearlin & Johnson, 1977). Persons who reported to have a symptom in the previous week were coded 1, all others were coded 0. Five symptoms were measured, namely, lack of enthusiasm, low energy, sleeping problems, poor appetite, and feeling hopeless. The data set consisted of 748 males. Research has shown that a 2-class model does not adequately fit these data while a 3-class model does (Magidson & Vermunt, 2001). Here, we will check whether the same result is obtained using the new posterior-cppp.

First a latent class model was fitted with 2 latent classes. Model misfit was assessed using the Pearson χ^2-statistic to test if conditional independence was violated over all variables, as well as with the BVR to test if conditional independence was violated between pairs of variables. This was done using the standard ppp and the posterior-cppp. The results can be found in Table 3.7. As can be seen, the ppp for the χ^2 test is not significant using a significance level of .05 while the posterior-cppp is significant with a value of .001. Furthermore, none of the BVR-tests are significant using the ppp while the posterior-cppp suggests there is model misfit for variable pairs (1,2), (2,5), (3,4), (3,5), and (4,5) using a significance level of .05. Furthermore the BVR for the variable pair (2,5) is still significant after a Bonferroni correction of the significance level to .05/10 = .005. These results show that the standard ppp is unable to detect model misfit and would thus result in incorrect conclusions about the true number of latent classes. The posterior-cppp on the other hand does not have this problem.

Because of the misfit of the 2-class model as indicated by the posterior-

Table 3.7: Model fit results for the depression data

	2-class model		3-class model	
Discrepancy	ppp	posterior-cppp	ppp	posterior-cppp
χ^2	.140	.001	.489	.597
BVR_{12}	.424	.013	.415	.011
BVR_{13}	.549	.790	.545	.915
BVR_{14}	.498	.341	.523	.768
BVR_{15}	.500	.345	.557	.972
BVR_{23}	.512	.558	.545	.934
BVR_{24}	.502	.429	.509	.524
BVR_{25}	.258	.002	.389	.031
BVR_{34}	.143	.008	.314	.017
BVR_{35}	.159	.018	.521	.719
BVR_{45}	.239	.041	.527	.805

cppp, a 3-class model was fitted as well. The ppp's and posterior-cppp's can be found in Table 3.7. In this case, the posterior-cppp's do not indicate any serious form of model misfit. This analysis confirms that the depression scale data for males can be adequately described using a 3-class model.

3.6 Discussion

Posterior predictive checking is a very flexible methodology to evaluate various forms of model misfit without relying on large sample theory. The most commonly used posterior predictive check is based on the posterior predictive p value (ppp), which can efficiently be computed using MCMC output (Gelman et al., 1996; Meng, 1994). A problem with this approach however is that the ppp is generally too conservative, which results in tests with very low statistical power.

The prior-calibrated posterior predictive p value (prior-cppp) resolves this issue by calibrating the ppp under a proper prior distribution for all

model parameters. The prior-cppp simultaneously checks model misfit and prior misfit. As a result, the choice of the prior that is used for the calibration has a serious effect on the outcome of the test. Therefore the prior-cppp should only be used when clear prior information is available for all model parameters and one is interested in simultaneously testing the model and the prior. In our experience however this is hardly ever the case. Either clear prior information is not available for all model parameters, or researchers are only interested in testing whether the employed statistical model fits the observed data.

As an alternative the posterior-calibrated posterior predictive p value (posterior-cppp) was proposed. The posterior-cppp is obtained by calibrating the ppp under the posterior given the observed data under the null model. The posterior-cppp has all the advantages of the original ppp, i.e., it can be used to detect any form of misfit, it can be computed from MCMC output, and it can be computed using non-informative improper priors. In addition, the posterior-cppp also results in accurate type I error rates, which is not the case for the original ppp. Moreover, the posterior-cppp results in more statistical power than the ppp. The usefulness of the posterior-cppp was illustrated in different testing problems, such as testing independence of dichotomous variables, assessing misfit of regression models in the case of extreme observations, and testing misfit in latent class models.

A potential drawback of the posterior-cppp (and the prior-cppp) is that it requires more computational time than the standard ppp because an additional calibration step is needed. In order to compute the posterior-cppp, say, 500 ppp's need to be calculated which takes maximally 500 times longer than computing the standard ppp. These 500 ppp's however can be computed in parallel and therefore computation time can be drastically

reduced. In the empirical regression application, for example, the computation of the standard ppp was .05 seconds and the computation of the posterior-cppp was 7.1 seconds. For this reason we believe the additional computational time of the posterior-cppp hardly limits its applicability.

3.A Deriving the Posterior when Testing for Independence

For a data set with sample size N, let n_{1j} and $N - n_{1j}$ be the observed number of persons with scores 1 and 0 on the j-th variable, and let π_{1j} and $1-\pi_{1j}$ be the corresponding probabilities, for $j = 1, \ldots, J$. Thus, the vector of model parameters consists of the unknown probabilities $(\pi_{11}, \ldots, \pi_{1J})$. Under the assumption that the J variables are independent, the likelihood is obtained as a product of J independent binomial distributions, i.e.,

$$p(n_{11}, \ldots, n_{1J}|\pi_{11}, \ldots, \pi_{1J}) = \prod_{j=1}^{J} p(n_{1j}|\pi_{1j}) \tag{3.22}$$

where the term for the j-th variable is proportional to

$$p(n_{1j}|\pi_{1j}) \propto (\pi_{1j})^{n_{1j}}(1 - \pi_{1j})^{N-n_{1j}}. \tag{3.23}$$

The beta distribution is the conjugate prior for the binomial model, i.e.,

$$\begin{aligned} p(\pi_{1j}) &= Beta(\pi_{1j}|\alpha_j, \beta_j) \\ &\propto (\pi_{1j})^{\alpha_j - 1}(1 - \pi_{1j})^{\beta_j - 1}. \end{aligned}$$

The hyper-parameters α_j and β_j can be specified in accordance with our prior knowledge (or lack thereof) regarding the distribution of the response

probability of variable j, for $j = 1, \ldots, J$. Note that when $\alpha_j = \beta_j = 1$, a uniform prior is obtained for π_{1j}. Multiplying the prior and the likelihood according to (3.1), yields the posterior for π_{1j} which is given by

$$\begin{aligned} p(\pi_{1j}|n_{1j}) &\propto p(n_{1j}|\pi_{1j}) \times p(\pi_{1j}) \\ &\propto (\pi_{1j})^{n_{1j}+\alpha_j-1}(1-\pi_{1j})^{N-n_{1j}+\beta_j-1} \\ &\propto Beta(\pi_{1j}|n_{1j}+\alpha_j, N-n_{1j}+\beta). \end{aligned}$$

3.B Posterior Distributions for the Regression Model Parameters

We use the standard independence Jeffreys prior $\pi(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$ throughout our regression analyses (e.g., Kass & Wasserman, 1996). The (conditional) posteriors used in Step 2 of Algorithm 1 are then given by

$$\begin{aligned} \sigma^2|\mathbf{y}, \mathbf{X} &\sim IG(\tfrac{n-J}{2}, \tfrac{s^2_{\mathbf{y}}}{2}) \\ \boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X} &\sim N(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}), \end{aligned}$$

where the ML estimate is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, the sum of squares equals $s^2_{\mathbf{y}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$, and $IG(\alpha, \gamma)$ denotes an inverse gamma distribution with shape parameter α and scale parameter γ. The annotated Julia (Bezanson, Edelman, Karpinski, & Shah, 2014) code used for the computation of the posterior-cppp was added as supplemental material to the publication and can be provided upon request.

3.C Latent Class Analysis Technical Details

For a data set with sample size N, suppose there are C latent classes. Let ν_c be the (unobserved) class size of the latent class c and let n_{1jc} and $\nu_c - n_{1jc}$ be the observed number of persons with scores 1 and 2 on the j-th variable within class c. Then let ρ_c be the class proportions, and let π_{1jc} and $1 - \pi_{1jc}$ be the corresponding conditional response probabilities, for variables $j = 1, \ldots, J$ and classes $c = 1, \ldots, C$. Thus, the vector of model parameters $\boldsymbol{\theta}$ consists of the class proportions $(\rho_1, \ldots, \rho_C)$ and the unknown conditional response probabilities $(\pi_{11c}, \ldots, \pi_{1Jc})$. Under the main latent class model assumption that the J variables are *conditionally* independent, the likelihood is obtained as a weighted sum of the product of J conditionally independent binomial distributions, i.e.,

$$p(n_{11c}, \ldots, n_{1Jc} | \pi_{11c}, \ldots, \pi_{1Jc}) = \sum_{c=1}^{C} \rho_c \prod_{j=1}^{J} p(n_{1jc} | \pi_{1jc}) \tag{3.24}$$

where the term for the j-th variable is proportional to

$$p(n_{1jc} | \pi_{1jc}) \propto (\pi_{1jc})^{n_{1jc}} (1 - \pi_{1jc})^{\nu_c - n_{1jc}}. \tag{3.25}$$

Again, it is standard practice to use the conjugate $Beta(\pi_{1jc} | \alpha_j c, \beta_j c)$ priors for the conditional response probabilities. Since the number of classes can be greater than two, we assume a multinomial (rather than a binomial) likelihood for the class proportions. The conjugate prior for the multinomial likelihood is the multivariate generalisation of the Beta distribution given as $Dirichlet(\rho_c | \alpha_1, \ldots, \alpha_C)$.

By combining the likelihood and the prior using (3.1), the posterior distribution for π_{1jc} used in Step 2 of Algorithm 1 has a beta distribution

given by $p(\pi_{1jc}|y_j) = Beta(\pi_{1jc}|n_{1jc} + \alpha_{jc} - 1, \nu_c - n_{1jc} + \beta_{jc} - 1)$ and the posterior for the vector of class proportions is given by $p(\rho_1, \ldots \rho_C|\mathbf{y}) = \text{Dirichlet}(\rho_1, \ldots, \rho_C|\nu_1 + \alpha_1 - 1, \ldots, \nu_C + \beta_C - 1)$. These posteriors assume that it is known which observations belong to which class, in order to obtain the observed frequencies ν_c and n_{rjc}. In order to do this, the data needs to be augmented with starting values for the latent class memberships (Tanner & Wong, 1987), after which a Gibbs sampler is run to iteratively draw values from the posteriors and then updating the latent class memberships. When the Gibbs sampler has converged (usually after thousands of iterations), values for the parameters are obtained by retaining every 50th, or so, draw.

3.C.1 Calculating the bivariate residual.

Cross-tabulating two dichotomous variables, results in a contingency table with $2^2 = 4$ cells. Each cell corresponds to one of the four response patterns, denoted by $\mathbf{y}_s$, for $s = 1, \ldots, 4$. with pattern probability denoted by π_s and given by

$$\pi_s = \sum_{c=1}^{C} \rho_c (\pi_{1jc})^{d_{js}} (1 - \pi_{1jc})^{1-d_{js}} \times (\pi_{1j'c})^{d_{j's}} (1 - \pi_{1j'c})^{1-d_{j's}}, \tag{3.26}$$

where the dummy indicator d_{js} and $d_{j's}$ equal 1 if the response to variable j and j' in pattern s are 1, and 0 otherwise. For example the response probability for the pattern (1,1) equals $\sum_{c=1}^{C} \rho_c \pi_{11c} \pi_{12c}$. For example, the expectations used to calculate in the D_{BVR} for pattern (1,1), given the model parameters at iteration k would equals $e_s^{(k)} = N \sum_{c=1}^{C} \rho_c^{(k)} \pi_{11c}^{(k)} \pi_{12c}^{(k)}$. Annotated Julia code for calculating a posterior-cppp for the BVRs was added as supplemental material to the publication and can be provided

upon request.

3.C.2 Calculating the Pearson χ^2 and the likelihood ratio.

Cross-tabulating all J dichotomous variables, results in a contingency table with $S = 2^J$ cells. Each cell corresponds to one of the possible response patterns (which now comprise J responses), again denoted by $\mathbf{y}_s$, for $s = 1, \ldots, S$. The pattern probability is denoted by π_s and given by

$$\pi_s = \sum_{c=1}^{C} \rho_c \prod_{j=1}^{J} (\pi_{1jc})^{d_{js}} (1 - \pi_{1jc})^{1-d_{js}}, \tag{3.27}$$

where the dummy indicator d_{js} equals 1 if the response to variable j in pattern s is 1, and 0 otherwise. For example the response probability for only ones on all J variables equals $\sum_{c=1}^{C} \rho_c \pi_{11c} \times \ldots \times \pi_{1Jc}$ (which may differ across iterations of the Gibbs sampler).

Chapter 4

Fast Resampling Method for Evaluating Latent Class Model Fit

Abstract

The latent class model is a powerful unsupervised clustering algorithm for categorical data. Many statistics exist to test the fit of the latent class model. However, traditional methods to evaluate those fit statistics are not always useful. Asymptotic distributions are not always known, and empirical reference distributions can be very time consuming to obtain. In this chapter we propose a fast resampling scheme with which any type of model fit can be assessed. The principle behind the method is to specify a statistic which captures the characteristics of the data that a model should capture correctly. If those characteristics in the observed data and in model-generated data are very different we can assume that the model could not have produced the observed data. With this method we achieve the flexibility of tests from the Bayesian framework, while only needing maximum likelihood estimates. We provide a step-wise algorithm with which the fit of a model can be assessed based on the characteristics we as researcher find

This chapter is currently being prepared for submission

important. In a Monte Carlo study we show that the method has very low type I errors, for all illustrated statistics. Power to reject a model depended largely on the type of statistic that was used and on sample size. We applied the method to an empirical data set on clinical subgroups with risk of Myocardial infarction and compared the results directly to the parametric bootstrap. The results of our method were highly similar to those obtained by the parametric bootstrap, while the required computations differed three orders of magnitude in favour of our method.

4.1 Introduction

The latent class (LC) model (Goodman, 1974) is a powerful unsupervised clustering algorithm for categorical data that is currently being used in a wide range of research fields. An important part of doing LC analysis is to assess the fit of the model to the observed data. Besides the traditional χ-squared goodness-of-fit statistics, various specific statistics have recently been developed for this model and its extensions (Vermunt & Magidson, 2016; Oberski et al., 2013; Nagelkerke, Oberski, & Vermunt, 2016a, 2016b).

Deciding on whether the value of a statistic indicates model misfit is usually based on a p value, which quantifies how likely the observed data are if the employed model holds in the population. The asymptotic p value is the most commonly used p value. It is very easy to calculate, but can only be obtained if the asymptotic distribution of a statistic is known, which is not the case for all available statistics (van Kollenburg et al., 2015; Oberski et al., 2013). Additionally, the asymptotic approximation may break down when sample sizes are not large so that the resulting p values can become very unreliable (van Kollenburg et al., 2015). More reliable p values can be obtained by using computational intensive resampling schemes yielding empirical reference distributions. This is most commonly done by means

of the parametric bootstrap (Efron & Tibshirani, 1993). The downside of the bootstrap method is that it can become very time-consuming when the models under investigation are complex, such as the LC model, because the same model has to be estimated many times (Nagelkerke et al., 2016b). And even then, certain statistics may still result in unreliable p values (Von Davier, 1997; Maydeu-Olivares & Joe, 2006; van Kollenburg et al., 2015; Oberski et al., 2013); for example, bootstrap p values for overall goodness-of-fit statistics are unreliable when the number of possible data patterns is very large. Efforts to speed up the bootstrap procedure usually focus on parameter estimation (Antal & Tillé, 2014; Salibian-Barrera & Zamar, 2002) or are limited to small data sets (Kisielinska, 2013). In the Bayesian framework, there is a tool available to test model fit, called the posterior predictive check (Meng, 1994; Gelman et al., 1996). Though it is fast and does not require repeated model estimation, the need for MCMC algorithms and choices for prior distributions may be a barrier for applied researchers.

Ideally, a test for model fit should be fast, reliable, and easy to implement. The methodology proposed in this chapter improves on existing methods by eliminating the need for repeated model estimation, while still only requiring maximum likelihood estimates. The idea is to directly compare observed data with model-generated data in a way that requires no time-consuming model estimation procedures. This comparison can be based on characteristics corresponding to those aspects of the data that the researcher finds most important. If the observed data differ consistently from the model-generated data, we have strong evidence that the model under consideration could not have produced the observed data. The proposed methodology cannot only be used to assess how well a model can explain particular aspects of the data, it can also be used to assess whether

the key model assumptions are met when applying the model concerned on the data set at hand. As we will show, the proposed methodology provides great flexibility to assess many different aspects of model fit.

As an example, suppose we are interested in testing whether a latent LC model is appropriate to describe a data set based on a risk inventory. For each observation, a number of variables are scored according to whether a particular risk is present (scored 1) or not present (scored 0). A first research question would be how many latent subgroups are needed to capture the associations between those variables. Second, we may require the model to correctly capture the observed frequency of observations who have all risk present. For the first research question, it is straightforward to quantify the association by a statistic such as the Pearson's χ^2 statistic under independence. For the second question, the statistic of interest is simply the observed frequency that this particular pattern occurred. Suppose we have an observed $\chi^2 = 200$ and that, out of 100 observations, the pattern with all risks present occurs 25 times. We then generate many data sets from the estimated latent class model. If in those generated data sets we find χ^2-values around 200 and frequencies of the response pattern of interest of around 25, the employed model seems valid to explain the characteristics of the observed data. The reason is that the data generated from the fitted LC model has similar characteristics as the observed data.

In this chapter, we shall focus on tests to assess whether a LC model adequately captures specific characteristics of the data. We will achieve this by using test statistics which capture characteristics of a data set and which can be calculated directly from the data (e.g., descriptive statistics) without needing to re-estimate the model for each replicated data set using time-consuming algorithms such as the Expectation Maximization (EM) algorithm. Although the proposed procedure is similar to the

parametric bootstrap (Efron & Tibshirani, 1993), the key difference is that our methodology directly compares observed and replicated data, rather than in an indirect manner as is done when using a bootstrap of traditional goodness-of-fit statistics. It should be noted that the parametric bootstrap needs to estimate the model of interest on both the observed data and the (large number of) generated data sets because a goodness-of-fit statistic needs model estimates for its calculation. Due to the indirect nature of testing with the parametric bootstrap, our method has the advantage that it can be up to hundreds of times faster, because the model only needs to be estimated once rather than hundreds of times as is required for the parametric bootstrap. The second advantage of using tailored test statistics is that we obtain a detailed picture of which aspects of the data are and which aspects are not captured by the model. In contrast to the posterior predictive check (Meng, 1994; Gelman et al., 1996), the proposed method is based on maximum likelihood estimation, which is generally easier to perform than Bayesian MCMC sampling and does not require specification of prior distributions.

In the remainder of this chapter we will illustrate the proposed resampling methodology within the context of LC analysis. In the next section we discuss the LC model in detail and provide statistics with which the fit of a LC model can be tested. After that we introduce the new methodology in more detail, and provide a step-wise algorithm for applying the method to LC analysis. We include a simulation study evaluating type I error rates and power for various statistics of interest. The chapter ends with a discussion on the main finding, their implications for applied researchers, and interesting topics for future research.

4.2 The Latent Class Model

The LC model (Goodman, 1974) is used to cluster observations when there is no knowledge about which observation belongs to which group (i.e., it is unsupervised). Suppose we have N observations on J dichotomous variables (e.g., yes/no, present/absent, male/female). Each observation can have one of $S = 2^J$ possible response patterns $\boldsymbol{y}_s = (y_{s1}, \ldots, y_{sJ})$, for $s = 1, \ldots, S$. The LC model assumes that the N observations can be divided into C distinct classes/categories of an unobserved variable based on their response pattern.

According to the LC model, each class has a different probability of giving a particular response to the variables. We denote the probability of giving response 1 to variable j given that an observation belongs to class c as π_{jc}. The (unknown) class sizes are denoted by ρ_c. The LC model further assumes that within a class, the responses to the variables are independent (i.e., local independence). Therefore, we can write the LC model likelihood of observing a particular response pattern as

$$Pr(\boldsymbol{y}_s) = \sum_{c=1}^{C} \rho_c \prod_{j=1}^{J} \pi_{jc}^{y_{sj}} (1 - \pi_{jc})^{1-y_{sj}}. \tag{4.1}$$

As an example, the response probability for a pattern with only ones on all J variables is equal to $\sum_{c=1}^{C} \rho_c \times \pi_{1c} \times \ldots \times \pi_{Jc}$. Maximum likelihood estimates for the model parameters are usually obtained though the Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977).

4.2.1 Statistics for the latent class model.

The focus of LC analysis often lies with explaining associations between variables. A straightforward choice to assess the strength of the association between categorical variables is to calculate a chi-squared type of statistic (Agresti & Yang, 1987) quantifying the deviation from independence. Commonly used measures are Pearson's X^2,

$$X^2 = \sum_{s=1}^{S} \frac{(n_s - e_s)^2}{e_s},$$

and the likelihood ratio chi-square,

$$G^2 = 2 \sum_{s=1}^{S} n_s \ln(n_s/e_s),$$

where n_s is the observed frequency of a particular pattern $\boldsymbol{y}_s$ and e_s is the expected frequency of that pattern under independence. It is important to note that the e_s can easily be computed from the data, that is, by multiplying the univariate marginal frequencies of the variables involved. These chi-squared type of statistics can be used to quantify associations of any order, from second-order for bivariate associations to J-th order for the overall association between all J variables used in the analysis.

Another example of a question that a researcher may be interested in is whether we can find an explanation for the occurrence of particular response patterns. For instance, suppose we have data on some sort of risk inventory. Each dichotomous variable in the inventory indicates whether a particular risk is present (scored 1) or not present (scored 0). Now suppose that we focus on the number of high-risk observations for which at least a certain number (say, Q) of the risks are present. The statistic used to

test this may simply be the total frequency of the patterns with Q or more risks present:

$$D_{\text{risk}}(\mathbf{y}, R) = \sum_{s=1}^{S} n_s * I(\sum_{j=1}^{J} y_{sj} \geqslant Q)$$

where n_s is the observed frequency of pattern $\boldsymbol{y}_s$ and where the indicator function I equals 1 if the sum of the scores in pattern s is greater than or equal to Q (i.e., it has Q or more risks present) and 0 otherwise. The formula then sums all the corresponding frequencies, n_s. Aside from risk assessment, other examples include presence of symptoms, giving correct answers, and agreement to statements. The statistic can easily be adapted to assess whether the frequency of one particular response pattern (e.g., $[1, 0, 0, 1, 0, 1]$) is correctly picked up. In that case, the statistic is reduced to simply the observed frequency n_s for the pattern of interest.

The LC model is of particular interest for the above stated research questions, because it may explain that the associations found in the data are a result of the fact that the data is a combination of data from different subgroups. Additionally, the LC model might explain that the observed frequency of risk patterns is due to the fact that there is a specific subgroup which is prone to have high risks, while another subgroup has not.

It is important to note that the chi-squared statistics that we use in the current application are traditionally used within LC modelling as residuals to test *remaining* association given a particular model. The fit of a LC model is then usually evaluated through the parametric bootstrap (Oberski et al., 2013; van Kollenburg et al., 2015; Nagelkerke et al., 2016b). When used as residuals, the expected frequencies e_s are based on the ML estimates (plugged into Equation 4.1). The implication is that ML estimates also have to be found for every replicate data set, which due to the need for iterative estimation algorithms may make the parametric bootstrap very

time consuming. The field can therefore greatly benefit from the use of tailor-made statistics which can be calculated directly on the data.

4.3 A New Methodology to Test Model Fit

The general idea of resampling methods is to obtain an approximation of the distribution of a statistic without the requirement of relying on asymptotics. To assess model fit, we check whether a model of interest has a similar fit to the observed data (quantified by a goodness-of-fit statistic) as it has to (hypothetical) model-generated data (quantified by the same statistic). When the fit in replicated data and observed data are similar, this suggests that the model fits the data well. When the fit in the observed data is much worse than in the generated data, this suggests that the model does not fit. The underlying principle is that if the observed data was generated by the model, it should be similar to artificial data of which we know that it was generated from the model.

The methodology proposed here applies the same basic principle of all resampling methods, that is, by quantifying important characteristics of the data and comparing those characteristics in observed and model-generated data. The following algorithm describes how to assess the fit of a LC model based on a statistic calculated directly on the data. The algorithm can be applied to any statistic and any model specification. For illustration purposes we will include in the description how to use the X^2 for testing the fit of a 2-class model.

Algorithm 1: Comparing characteristics of observed and model-generated data

Step 1: Specify the important characteristics of the observed data and calculate the corresponding statistic.

We are interested in correctly reproducing the overall association between the variables as quantified by

$$X^2 = \sum_{s=1}^{S} \frac{(n_s^{\text{obs}} - e_s^{\text{obs}})^2}{e_s^{\text{obs}}}.$$

Step 2: Specify the likelihood of the model and obtain ML estimates for the model parameters.

The likelihood for a particular response in a 2-class model is given by

$$Pr(\boldsymbol{y}_s) = \sum_{c=1}^{2} \rho_c \prod_{j=1}^{J} \pi_{jc}^{y_{sj}} (1 - \pi_{jc})^{1-y_{sj}}, \tag{4.2}$$

where ML estimates for the parameters can be obtained through the EM algorithm.

Step 3: Obtain values for the chosen statistics in replicated data sets:

Step 3a: Plug in the ML estimates into the likelihood specified in Step 2.

We calculate the probability for each response pattern, $\boldsymbol{y}_s$, by plugging the ML estimates into the likelihood:

$$\widehat{Pr}(\boldsymbol{y}_s) = \hat{\rho}_1 \prod_{j=1}^{J} \hat{\pi}_{j1}^{y_{sj}} (1 - \hat{\pi}_{j1})^{1-y_{sj}} + \hat{\rho}_2 \prod_{j=1}^{J} \hat{\pi}_{j2}^{y_{sj}} (1 - \hat{\pi}_{j2})^{1-y_{sj}}.$$

Step 3b: Generate a replicated data set, $\mathbf{y}_{\text{rep}}^{(k)}$ from the likelihood with the ML estimates plugged in.

A replicate data set is generated by taking a draw from a multinomial distribution, with the estimated pattern probabilities as parameters:

$$\mathbf{y}_{\text{rep}}^{(k)} \sim \text{Multinomial}(N|\widehat{Pr}(\mathbf{y}_1), \ldots \widehat{Pr}(\mathbf{y}_S)).$$

Step 3c: Calculate the chosen statistic on the replicated data set.

We compute the association between all the variables in a replicate data set $\mathbf{y}_{\text{rep}}^{(k)}$ as:

$$X^{2,(k)} = \sum_{s=1}^{S} \frac{(n_s^{(\text{k})} - e_s)^2}{e_s^{(\text{k})}}.$$

where the expected frequency e_s is computed from the likelihood as the product of the corresponding marginal probabilities multiplied by the sample size.

Step 3d: Repeat Steps 3b and 3c for $k = (1, \ldots, K)$, for large K. Figure 4.1 depicts the resulting distribution of $K = 1000$ values of the statistic based on 1000 replicated data.

Step 4: Compute the proportion of replicated data sets where the value for the statistic was greater than or equal to the value for the observed statistic.

$$\text{p}(X^2) = K^{-1} \sum_{k=1}^{K} I(X^{2,(k)} \geqslant X^2)$$

where the indicator function $I(\cdot)$ equals 1 if the value of the statistic in the replicate data is larger than, or equal to the value of the statistic in the observed data.

In summary, the goal of the methodology is to obtain the sampling

distribution of the statistic of interest using a large number of replicated data sets for which the statistic of interest is computed. This results in K values for the statistic, which are then compared to the value of the statistic in the observed data. As an example Figure 4.1 displays the values for the Pearson's X^2 calculated on $K = 500$ data sets generated from a two-class model which was estimated on an empirical data set (which will be discusses later on). The p value is equal to the proportion of generated data sets in which the value for the X^2 was at least as large as the observed X^2. Visually, this can be represented as the vertical line in Figure 4.1 at the value of $X^2 = 226.236$. The p value can then be seen as the area under the curve to the right side of that line. In this example, the p value was equal to 0.266.

Note that we can specify multiple statistics simultaneously in Steps 1 and 3c of the Algorithm. It is important to realise that the implication of a small p value depends on the corresponding statistic. For example, a small p value for the X^2 means that the variables in the model-generated data contain less association than in the observed data (only few $X^{2,(k)}$ values were as extreme as the observed X^2). The conclusion is then that the given model does not appropriately explain the associations in the observed data. However, other statistics may have a reverse interpretation, where high p values indicate model misfit. For instance, it may be crucial for an application that a model does neither over- nor underestimate the occurrence of an important response pattern. In such a case, both a high and a low p value for the frequency n_s indicates model misfit.

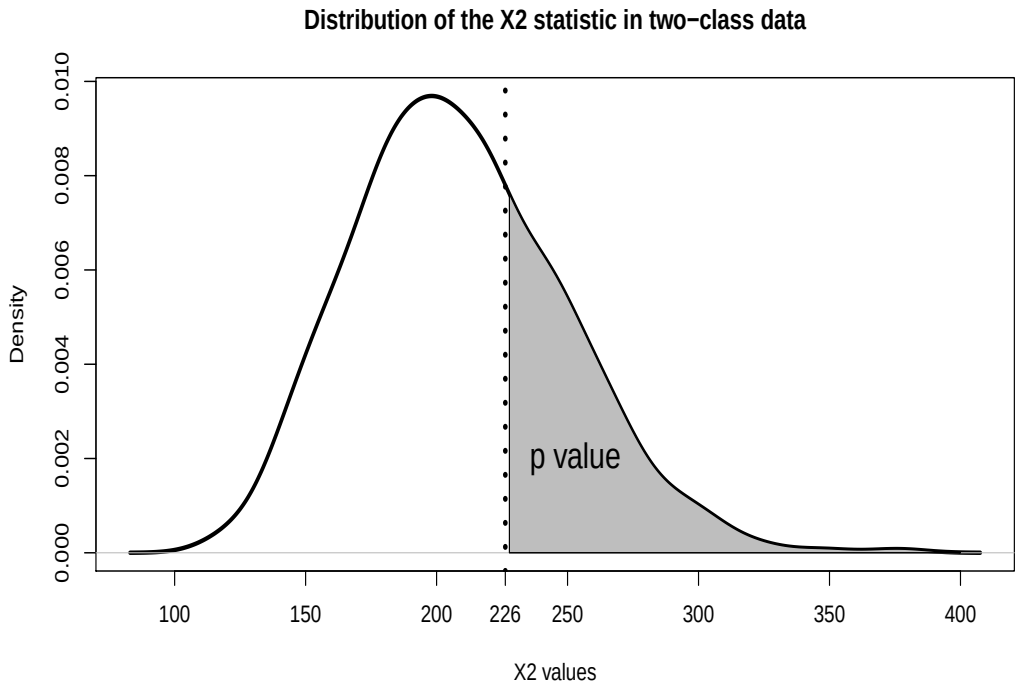

Figure 4.1: The distribution of the X^2 statistic calculated on 500 data sets generated from a LC model with 2 classes. The value of X^2 in the observed data is represented by a vertical dotted line. The proportion of the area under the curve to the right of the vertical line (i.e., .296) is used to make the decision about the appropriateness of the model.

4.3.1 Simulation study.

We set up a Monte Carlo study to assess type I error rates and power of a number of statistics for evaluating the fit of a LC model. For this, we generated data on $J = 6$ dichotomous variables from latent class models with either $C = 2$ or $C = 3$ classes. Each class was of equal size (i.e. $\rho_c = 1/C$, for all c). Conditional probabilities for a 1 response were $\pi_{jc} = .8$ (or .9) in class 1 and $\pi_{jc} = .2$ (or .1) in class 2 for all $J = 6$ dichotomous variables. For class 3, the conditional probabilities for the 1 response were .8 (.9) for the first three variables and .2 (.1) for the last three variables. The parameter values for the conditions for a three-class model with $\pi_{j1} =$

Table 4.1: Class proportions ρ_c and conditional response probabilities $\pi_1 c$ for the simulation conditions with $C = 3$ classes.

class	$c = 1$	$c = 2$	$c = 3$
ρ_c	1/3	1/3	1/3
π_{1c}	.8	.2	.8
π_{2c}	.8	.2	.8
π_{3c}	.8	.2	.8
π_{4c}	.8	.2	.2
π_{5c}	.8	.2	.2
π_{6c}	.8	.2	.2

.8 are shown in Table 4.1. We used sample sizes of $N = 100, 500$ and 1000. For each combination of conditions we generated 1000 data sets.

For each generated data set we evaluated whether the LC model with 2 classes was able to reproduce the total association between all variables (quantified by the Pearson's χ^2 and the likelihood ratio G^2), the bivariate association between the first two variables (quantified as χ^2_{12}), and the occurrence of the highest-risk pattern with all variables scored 1 (quantified as $D_{\text{risk}}(\mathbf{y}, 6)$). Data was generated using R 3.3 (R Core Team, 2016) and subsequent analyses were done with LatentGOLD 5.1 (Vermunt & Magidson, 2016). The results for the type I error rates and power can be found in Table 4.2.

The first 6 rows of Table 4.2 show the type I error rates, which are all very low. That is, regardless of the chosen statistic we rarely rejected the 2-class model as being the data generating process when this was indeed the case. From the last 6 rows, we can see how often we rejected the 2-class model when the 'observed' data was generated from a 3-class model. The power to reject a 2-class model based on aggregated associations in the data was very high (columns 4 and 5), except when sample size was small ($N = 100$) and response probabilities corresponded to $\pi_j 1 = .8$. The power to

Table 4.2: Estimated type-I error rates (rows where $C = 2$) and power (rows where $C = 3$) when testing the appropriateness of a 2-class LC model when using a significance level of .05. Risk(6) indicates risk statistic for the pattern with all risks present.

C	N	π_{j1}	χ^2	G^2	χ^2_{12}	$Risk(6)$
2	100	.8	$.002 \pm .002$	$.006 \pm .003$	$.022 \pm .007$	$.018 \pm .006$
		.9	$.014 \pm .005$	$.006 \pm .003$	$.012 \pm .005$	$.008 \pm .004$
	500	.8	$.000 \pm .000$	$.000 \pm .000$	$.002 \pm .002$	$.002 \pm .002$
		.9	$.000 \pm .000$	$.000 \pm .000$	$.004 \pm .003$	$.000 \pm .000$
	1000	.8	$.000 \pm .000$	$.002 \pm .002$	$.006 \pm .003$	$.002 \pm .002$
		.9	$.000 \pm .000$	$.000 \pm .000$	$.000 \pm .000$	$.002 \pm .002$
3	100	.8	$.180 \pm .017$	$.214 \pm .018$	$.292 \pm .020$	$.234 \pm .019$
		.9	$.990 \pm .004$	$.956 \pm .009$	$.556 \pm .022$	$.544 \pm .022$
	500	.8	$.934 \pm .011$	$.906 \pm .013$	$.648 \pm .021$	$.500 \pm .022$
		.9	$1.000 \pm .000$	$1.000 \pm .000$	$.544 \pm .022$	$.506 \pm .022$
	1000	.8	$1.000 \pm .000$	$.993 \pm .004$	$.807 \pm .018$	$.593 \pm .022$
		.9	$1.000 \pm .000$	$1.000 \pm .000$	$.659 \pm .021$	$.545 \pm .022$

reject the three-class model based on a single bivariate association (column 6) increased overall with bigger sample sizes, but for larger samples, more extreme response probabilities provided lower power. The risk statistic had a power of over .5, except for the smallest sample size and response probabilities of .8.

4.3.2 Application to Empirical Data

Data from Rindskopf and Rindskopf (1986) was used for an empirical illustration. It contains $J = 4$ dichotomous variables indicating risk factors of myocardial infarction (MI); that is, presence of Qwave in ECG, presence of flipped LDH, presence of CPK-MB, and presence of classical clinical history. Characteristics of the data were assessed by overall X^2 and G^2 statistics and bivariate X^2s. All total scores were evaluated as well (i.e.,

Risk statistic with 1, 2, 3, or 4 risks present).

We estimated a 1-class model and a 2-class model, and used our methodology to test the fit of both models. We also compared our methodology with the standard parametric bootstrap. In both methods, the resulting p values were calculated based on $K = 1000$ replicated data sets. The results in Table 4.3 indicate that the 1-class model does not capture the associations between the variables nor reproduces the observed frequencies of the risks. All p values were 0 in our method as well as in the parametric bootstrap. For the 2-class model, we did not find any small p value, indicating that the model fits the data well. Our method provided the same conclusion as the computationally more intensive parametric bootstrap.

Table 4.3: Results of the LC Model Fit Test for the Myocardial Infarction Data.

		1-class		2-class	
Statistic	Value	p value	Bootstrap	p value	Bootstrap
X^2	226.236	.000	.000	.266	.308
G^2	149.468	.000	.000	.490	.381
X^2_{12}	44.082	.000	.000	.354	.213
X^2_{13}	39.339	.000	.000	.482	1.00
X^2_{14}	25.034	.000	.000	.472	.288
X^2_{23}	41.534	.000	.000	.323	.379
X^2_{24}	24.425	.000	.000	.379	.584
X^2_{34}	25.824	.000	.000	.290	.225
Risk(1)	61	.000	NA	.543	NA
Risk(2)	46	.000	NA	.367	NA
Risk(3)	36	.000	NA	.633	NA
Risk(4)	24	.000	NA	.231	NA

The substantive interpretation of parameters of the 2-class model is straightforward. We encountered a class with low chance of all 4 risks (Class 1 is the group without MI) and a high risk class (Class 2 is the group with MI).

Table 4.4: Estimated parameter values of a two-class model for the myocardial infarction data. The values for π_{jc} give the conditional probabilities of a risk factor being present in that class.

		Class 1	Class 2
	ρ_c	.542	.458
Q-wave	$\pi_1 c$	.000	.767
LDH	$\pi_2 c$	.027	.828
CPK	$\pi_3 c$	.195	1.000
History	$\pi_4 c$	.195	.791

4.4 Discussion

In this chapter, we proposed a resampling scheme which combines the ideas of the Bayesian posterior predictive check (Meng, 1994; Gelman et al., 1996) with frequentist testing procedures, leading to a highly flexible and very fast test for statistical model fit. A detailed description of the methodology was given in the form of a stepwise algorithm. After that the methodology was applied in the context of latent class analysis, where type I error rates and power for different types of model fit tests were evaluated by means of a Monte Carlo study. An application to an empirical data set was provided to test the fit of a LC model used to assess clinical subtypes of patients with risk of myocardial infarction.

The conducted Monte Carlo study showed (very) low type I errors, which follows logically from the methodology. Low type I errors imply that model-generated data was similar to the observed data. This behaviour is also common for the posterior predictive check (e.g., Hjort et al., 2006; van Kollenburg et al., in press). Since the 'observed' data in the simulations was generated from the same model as the replicated data, these are expected to be similar. We found, unsurprisingly, that using a statistic which aggregates all information about the associations in the data had higher

power than statistics which used only bivariate associations. The results imply that data generated from a 3-class model has consistently different overall associations than data from a 2-class model. This also holds for bivariate associations, though the power is lower in that case. Interestingly, when sample size was smallest ($N = 100$), the bivariate χ^2_{12} had more power than the aggregated chi-squared statistics. This is likely due to sparseness of the full contingency table. The table had $S = 2^6 = 64$ cells with only 100 observations, making calculations of the association measures unreliable. The contingency table for χ^2_{12} only has 4 cells to fill with 100 observations and sparseness is not an issue. However, it was surprising that at higher sample sizes the response probability had a negative effect on the power of the χ^2_{12} statistic. Apparently, even with high average cell counts (i.e., when $N/4$ is large) extreme response probabilities may affect the reliability of the statistic. Whether this also holds for bivariate tests in traditional fit testing is an interesting topic for future research.

The results from our simulation can be directly compared to the parametric bootstrap and posterior predictive check by using results from van Kollenburg et al. (2015). Those authors compared, among other things, the performance of the parametric bootstrap and the posterior predictive check on type I error rates and power for various statistics in LC analysis. In most cases, the traditional (direct) comparison of model residuals with observed data had higher power. Higher power does come with higher type I error rates as well as significantly longer computation time. Note that in larger sample sizes of $N = 1000$ the power of all methods using global fit statistics was approximately 1 already. Future research may focus on comparing different resampling methods more in-depth on type I errors, power, and computation time.

By applying the method to an empirical data set for the standard LC

analysis, we found that a 1-class model produced consistently different data sets than the data set we observed, resulting in p values of .000 for all statistics. When we estimated a 2-class model, the data sets that were generated from that model were similar to the observed data. With all p values being between .238 and .538, none of the statistics indicated model misfit. We also performed a parametric bootstrap for this data set, which did not lead to different conclusions about the fit of the model.

All in all, the method proposed in this chapter is faster and more flexible than traditional resampling techniques. However, it does provide a more conservative test. There are situations in which conservativeness is a welcome attribute of a test. For instance in substantive research, selecting more classes than necessary in a LC analysis due to random noise in the data (i.e., overfitting) can be more problematic than missing a class (van den Bergh, Schmittmann, & Vermunt, in press). Moreover, the current methodology allows researchers to determine exactly which of the aspects they want a model to properly explain/reproduce. As long as the chosen model is able to pick up the characteristics that are of importance to the conducted research, we need not worry about every other aspect of the model fit. If there are multiple models with which a researcher wants to explain the data, each of these models can be tested with the same methodology.

In this chapter we applied our methodology to numerical statistics, and calculated p values to aid us in assessing model fit. In some situations it may also be useful to visualise the data. This is not uncommon in the Bayesian framework. A clear example can be found in Gelman et al. (2004) who even used it as a cover of their book. With minor adjustments to our methodology, such visual tests are now also available in the frequentist setting. The statistic in Step 1 and 3b is then a plot and Step 4 will be a

visual inspection rather than the computation of a p value.

On a final note, some researchers require that a test has nominal type I errors. In that case it may be possible to calibrate the resulting p value in much the same way as van Kollenburg, Mulder, and Vermunt (2017) calibrated the posterior predictive p value. Such a calibration may lead to nominal type I error rates and higher power for the test, but will again require more computation. At which levels of model complexity the computational burden of our method and the parametric bootstrap is the same and what the differences in power are, remain open research questions to answer in the future.

Summary

This dissertation focusses on testing model fit of the latent class (LC) model. This model is a powerful tool with which observations can be clustered based on their combination of responses to a number of categorical variables. The research leading to this dissertation focused on testing whether a particular model can correctly describe the observed data by means of p values. In the different chapters p values are compared, improved, and sped up, respectively.

In the second chapter, the frequentist properties (type I error rates and power) of a the most common p values were evaluated for many commonly used statistics in LC analysis. A simulation study was conducted which involved calculating different p values for LC models based on sparse and non-sparse contingency tables. It was found that the use of asymptotic p values resulted in very distorted type I error rates when contingency tables were sparse. The parametric bootstrap and model-based PPC performed much better in this regard than the asymptotic method. It was also found that a number of statistics which used second order marginals of the tables were not strongly influenced by sparseness. The power study suggested that nearly all combinations of p value and statistics can be used detect misfit when the number of LCs is misspecified. When sample sizes become very small, however, we should resort to the local fit measures.

The third chapter discusses the distribution of the posterior predictive

p value (ppp). The common interpretation of a p value, requires that it is uniformly distributed if the null model is true. Since this uniformity assumption is not true for the ppp, a posterior-calibrated ppp was proposed which is obtained by calibrating the ppp under the posterior given the observed data under the null model. The advantage of this posterior-cppp is that it has all the advantages of the original ppp, but results in nominal type I error rates and has a (much) higher power to detect model misfit. The benefits of the calibrated p value over the original ppp were demonstrated in a number of simulation studies and empirical applications with different testing problems for independence models, linear regression models and latent class models. All studies showed that the proposed posterior-calibrated ppp was uniformly distributed under the null and had much higher power to detect model misfit than the original ppp.

The methodology proposed in the fourth chapter improves on existing methods by eliminating the need for repeated model estimation, while still only requiring maximum likelihood estimates. In this sense it is a best-of-both-worlds approach by combining the flexibility of Bayesian tests with frequentist estimation procedures. The idea is to directly compare observed data with model-generated data in a way that requires no time-consuming model estimation procedures. This comparison can be based on specific characteristics corresponding to those aspects of the data (e.g., descriptive statistics) that the researcher finds most important. If the observed data differ consistently from the model-generated data, we have strong evidence that the model under consideration could not have produced the observed data. The conducted Monte Carlo study showed that the method is much faster than traditional resampling techniques, though it provides a more conservative test. Unsurprisingly, the more information about the data a statistic incorporates, the higher its power to indicate model misfit in case

of non-fitting models. An empirical example showed that the new method led the same conclusions as the parametric bootstrap. It is noteworthy to state that the new method was computationally three orders of magnitude faster than the bootstrap.

Dankwoord (Acknowledgements)

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

Het is een mooie reis geweest die tot dit proefschrift heeft geleid. Als ik terug denk aan hoe ik zover ben gekomen zijn er een aantal mensen die een belangrijke rol hebben gespeeld om dit te bereiken. Naast alle collega's die feedback hebben gegeven op mijn werk, of die me gezelschap hebben gehouden tijdens de pauzes, wil ik nog een aantal mensen in enigszins chronologische volgorde persoonlijk bedanken.

Ten eerste wil ik mijn ouders bedanken. Jullie rol loopt door alle momenten van mijn leven. Het is voor jullie altijd vanzelfsprekend geweest dat ik een opleiding zou volgen. Jullie steun tijdens mijn academische reis is enorm belangrijk geweest. De eerdere jaren vooral financieel, en in de laatste jaren praktisch met bijvoorbeeld het zorgen voor de kids. Zonder jullie was ik nooit zover gekomen.

Een toch wel cruciale rol voor mijn academische loopbaan is er geweest voor de coördinator van mijn middelbare school dhr. D'Elfant. Bedankt dat je me een half leven geleden niet van school hebt gestuurd, terwijl dat

waarschijnlijk wel geoorloofd was. Op veel vlakken heb ik mezelf sindsdien ontwikkeld en nu zie ik in hoe veel ik had kunnen verspelen in die periode. Terugkijkend op diezelfde middelbare school periode is me duidelijk geworden dat de academische wereld me op het lijf geschreven is. Waar ik me echt op kon focussen en waar altijd iets goeds uitkwam waren de werkstukken waar ik zelf het onderwerp kon kiezen. Daarbij heb ik geleerd dat ik geniet van het uitzoeken van moeilijke onderwerpen. Daarom dank ik in het bijzonder mw. Tielen voor het feit dat ik voor wiskunde een interactief werkstuk mocht maken over fractals, en dhr. Hermans voor het stimuleren van het schrijven van mijn profielwerkstuk getiteld "Wie is ik?".

Mijn eerste contact met statistiek was tijdens mijn bachelor psychologie in een cursus van Marcel Croon. Ook de cursus die me twee jaar later daadwerkelijk enthousiast maakte over statistiek werd gegeven door hem (en waarvan ik tijdens mijn PhD traject de werkcolleges mocht geven). De helderheid waarmee de stof werd uiteengezet maakte het voor mij mogelijk om mijn medestudenten gelijktijdig bijles te geven in principale componenten analyse en factor analyse, terwijl dat voor mij eigenlijk ook nog nieuw was. Nu, een aantal jaren, een gezamenlijk paper en een gedeelde kamer later weet ik dat jouw kennis en inzicht in de statistiek werkelijk jaren hebben gespaard van de levens van iedereen die met vragen bij je kwamen. Bedankt voor het delen van je kennis en ervaring.

Daarnaast heb ik veel te danken aan Luc van Baest. Via jouw cursus over regressie analyse en onze discussies over wiskundige vraagstukken ben ik geprikkeld om mezelf meer te verdiepen, dingen uit te zoeken en contact te zoeken met mensen die meer wisten over bepaalde onderwerpen. Ook dankzij jouw hulp bij het vinden van studenten om les te geven heb ik mezelf kunnen ontwikkelen tot waar ik nu ben.

Omdat ik een jaartje uit Tilburg weg was om in Noorwegen te studeren,

kon ik niet direct instromen in de Research Master in Tilburg. Andries van der Ark gaf me de mogelijkheid een werkstuk te maken om aan te tonen of ik genoeg statistisch inzicht had. Hopelijk heb ik met dit proefschrift laten zien dat je een goede keuze hebt gemaakt. Ik ben in ieder geval voor altijd dankbaar dat je een uitzondering voor me wilde maken, voor de begeleiding tijdens de ReMa, en voor de kennismaking met R.

Tijdens de master hoorde ik over Bayesiaanse statistiek. En omdat het niet standaard was raakte ik er zoals gewoonlijk in geïnteresseerd. Het toeval wilde (al geloof ik daar niet meer in) dat er een Bayesiaan in het departement was geland genaamd Joris Mulder. De stage die ik bij jou kon volgen, Joris, heeft een geweldige invloed gehad op mij en mijn werk. Dat je het vertrouwen in me had om als copromotor mee te werken waardeer ik enorm. Bedankt ook voor je inzet op de momenten dat er deadlines waren. Twee dingen die ik in het bijzonder van je geleerd heb zijn het zelf doorhakken van knopen ("Tja, dat is een keuze") en acceptatie dat dingen nou eenmaal niet altijd gaan zoals ik het in mijn hoofd heb ("Tja, het is wat het is"). Deze twee korte, maar academisch oh zo bruikbare, zinnetjes zal ik nog lang gebruiken en hopelijk doorgeven aan anderen.

In het tweede jaar van de master volgde ik het vak 'Categorical Data Analysis', welke werd gegeven door Jeroen Vermunt. De modellen die in deze cursus werden besproken waren allemaal interessant, en ik merkte al zeer snel dat ik hier aanleg voor had. Ik begreep de samenhang tussen de modellen en had veel vragen over de details die niet direct voor de cursus nodig waren. Dat ik veel oog had voor detail merkte ik ook op in de verschillende lezingen die er gegeven werden voor ons departement. Bij bijna elk praatje kon ik wel een goede vraag stellen over iets kleins wat me was opgevallen, of soms blijkbaar niet precies klopte. Ik heb mezelf steeds voorgehouden dat deze vragen er misschien wel toe geleid hebben

dat jij, Jeroen, me een geschikte kandidaat vond om een promotietraject in te gaan. Binnen je Vici project heb ik de vrijheid gekregen om precies dat te doen wat ik leuk vond en te leren waar ik behoefte aan had. Jouw rol als promotor heeft me zowel op academisch als op persoonlijk niveau veel geleerd.

Mijn paranimfen, congrescompagnons en werkmaten Mattis (de Derde) en Erwin (met die lange haren). De kale met die bril is vereerd dat jullie mee op het podium willen staan. Erwin, ik herinner me helder de eerste keer dat ik je iets hoorde zeggen. Dat was overigens al bij de introductiemiddag van de research master, tijdens een speech van Klaas Sijtsma. Het onderwerp dat je toen aankaartte heeft tot vele gesprekken geleid, die ons soms tot diep in de nacht leidden. Jij hebt me scherp gehouden en bent een van de weinigen met wie ik mijn opgedane kennis daarover kan delen. Ook al zijn onze ideeën niet te verenigen, liggen ze toch maar twee woorden uit elkaar. Mattis, je bent een goed mens. Ik heb genoten van onze gesprekken en de tijd dat we in onze eigen niche écht onderzoek aan het doen waren. De tijd die ik met je doorbracht heeft me op verschillende vlakken veel inzicht gegeven. Wat hebben we met z'n drieën gelachen, ook al verging het lachen Mattis soms weer als het wat later werd en de humor flauwer. Ik hoop dat jullie beiden houden en krijgen wat het ook mag zijn dat jullie gelukkig maakt. Onze herinneringen aan $\mathcal{C}\&\mathcal{E}$ zullen daar vast bij helpen. I am no expert on these matters, but thank you.

Als laatste wil ik mijn gezin bedanken. Jullie hebben er voor gezorgd dat ik door ging wanneer het zwaar was, en waren een welkome afleiding wanneer ik thuis was. Willeke, mijn lief. Er zijn weinig dingen die ik zeker weet, maar er is geen twijfel dat wij samen horen te zijn. Je waakte ervoor dat ik mezelf niet verloor in het werk en was er altijd om naar me te luisteren en me advies te geven. Jens, de jongen die alles wil weten en

alles goed wil doen. Ik zie veel van mezelf in je terug en ik hoop dat ik je honger naar kennis nog lang mag stillen. Ik kijk uit naar de dag dat jij mij dingen gaat leren. Isa, mijn lieve prinses die voor iedereen zorgt. Jouw creativiteit en inzicht in anderen is geweldig. De liefde die je me geeft zou ik nergens voor willen ruilen. Maud, de zachtaardige slimmerik. Wat ben je grappig en wat ben je wijs. Wat je ook zult worden, ik hoop dat je voor altijd in mijn leven bent. Willeke, Jens, Isa en Maud, jullie maken me compleet en I love you met heel mijn hart.

Bibliography

Agresti, A. (2002). *Categorical data analysis*. John Wiley & Sons. Hoboken, New Jersey.

Agresti, A., & Yang, M.-C. (1987). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics & Data Analysis*, *5*(1), 9–21.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected papers of hirotugu akaike* (pp. 199–213). Springer.

Antal, E., & Tillé, Y. (2014). A new resampling method for sampling designs without replacement: the doubled half bootstrap. *Computational Statistics*, *29*(5), 1345–1363.

Bayarri, M., & Berger, J. (2000). P values for composite null models. *Journal of the American Statistical Association*, *95*(452), 1127–1142.

Berger, J. (2006). The case for objective bayesian analysis. *Bayesian analysis*, *1*(3), 385–402.

Berkhof, J., Van Mechelen, I., & Gelman, A. (2003). A bayesian approach to the selection and testing of mixture models. *Statistica Sinica*, 423–442.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2014). Julia: A fresh approach to numerical computing. *CoRR*, *abs/1411.1607*. Retrieved from `http://arxiv.org/abs/1411.1607`

Chaves, R. L., Chakraborty, A., Benziger, D., & Tannenbaum, S. (2014). Clinical and pharmacokinetic considerations for the use of daptomycin in patients with staphylococcus aureus bacteraemia and severe renal impairment. *Journal of Antimicrobial Chemotherapy*, *69*(1), 200–210.

Choi, J., Hui, S. K., & Bell, D. R. (2010). Spatiotemporal analysis of imitation behavior across new buyers at an online grocery retailer. *Journal of Marketing Research*, *47*(1), 75–89.

Collins, L. M., Fidler, P. L., Wugalter, S. E., & Long, J. D. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research*, *28*(3), 375–389.

Cressie, N., & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 440–464.

Crow, S. J., Swanson, S. A., Peterson, C. B., Crosby, R. D., Wonderlich, S. A., & Mitchell, J. E. (2012). Latent class analysis of eating disorders: Relationship to mortality. *Journal of abnormal psychology*, *121*(1), 225.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, *39*(1), 1–38.

Dufour, M., Brunelle, N., & Roy, É. (2013). Are poker players all the same? latent class analysis. *Journal of Gambling Studies*, 1–14.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC. Boca Raton, FL.

Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Formann, A. K. (2003). Latent class model diagnostics:a review and some

proposals. *Computational statistics & data analysis*, *41*(3), 549–559.

Gelman, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from unform. *Electronic Journal of Statistics*, *7*, 2595–2602.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. Chapman & Hall/CRC.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–759.

Gobbens, R. J., Krans, A., & van Assen, M. A. (2015). Validation of an integral conceptual model of frailty in older residents of assisted living facilities. *Archives of gerontology and geriatrics*, *61*(3), 400–410.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231.

Haberman, S. J. (1979). *Analysis of qualitative data. volume 2: New developments*. Academic Press.

Haberman, S. J. (1988). A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *Journal of the American Statistical Association*, *83*(402), 555–560.

Hagenaars, J. A. (1988). Latent structure models with direct effects between indicators local dependence models. *Sociological Methods & Research*, *16*(3), 379–405.

Hjort, N., Dahl, F., & Steinbakk, G. (2006). Post-processing posterior predictive p values. *Journal of the American Statistical Association*, *101*(475), 1157–1174.

Hoijtink, H. (1998). Constrained latent class analysis using the gibbs sampler and posterior predictive p-values: Applications to educational testing. *Statistica Sinica*, *8*, 691–711.

Holmquist, N. D., McMahan, C. A., & Williams, O. D. (1968). Variability in classification of carcinoma in situ of the uterine cervix. *Obstetrical & Gynecological Survey*, *23*(6), 580–585.

Hoverd, W. J., & Sibley, C. G. (2013). Religion, deprivation and subjective wellbeing: Testing a religious buffering hypothesis. *International Journal of Wellbeing*, *3*(2).

Jansen, B. R. J., & van der Maas, H. L. J. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review*, *17*(3), 321–357.

Jeffreys, H. (1961). Theory of probability, harold jeffreys. *International series of monographs on physics.*.

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Kisielinska, J. (2013). The exact bootstrap method shown on the example of the mean and variance estimation. *Computational Statistics*, *28*(3), 1061–1077.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363–374.

Langeheine, R., Pannekoek, J., & Van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research*, *24*(4), 492–516.

Lanza, S. T., Flaherty, B. P., & Collins, L. M. (2004). Latent class analysis and latent transition analysis. In J. Schinka & W. Velicer (Eds.), *Handbook of psychology: Volume 2. research methods in psychology (pp. 663–685).* Hoboken, NJ: Wiley.

Laudy, O., Zoccolillo, M., Baillargeon, R. H., Boom, J., Tremblay, R. E., & Hoijtink, H. (2005). Applications of confirmatory latent class analysis

in developmental psychology. *European Journal of Developmental Psychology*, *2*(1), 1–15.

Liang, F., Liu, C., & Carroll, R. (2011). *Advanced markov chain monte carlo methods: learning from past samples* (Vol. 714). John Wiley & Sons.

Ligtvoet, R., & Vermunt, J. K. (2012). Latent class models for testing monotonicity and invariant item ordering for polytomous items. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 237–250.

Lin, H., McCulloch, C. E., Turnbull, B. W., Slate, E. H., & Clark, L. C. (2000). A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations. *Statistics in Medicine*, *19*(10), 1303–1318.

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological methodology*, *31*(1), 223–264.

Magidson, J., & Vermunt, J. K. (2004). Latent class models. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences (pp. 175–198).* Thousand Oaks, CA: Sage Publications.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732.

Meng, X. L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*(3), 1142–1160.

Meulders, M., De Boeck, P., Kuppens, P., & Van Mechelen, I. (2002). Constrained latent class analysis of three-way three-mode data. *Journal of classification*, *19*(2), 277–302.

Mulder, J. (2014). Bayes factors for testing inequality constrained hypothe-

ses: Issues with prior specification. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 153–171.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*(1), 190–204.

Nagelkerke, E., Oberski, D. L., & Vermunt, J. K. (2016a). Goodness-of-fit of multilevel latent class models for categorical data. *Sociological Methodology*, *46*(1), 252–282.

Nagelkerke, E., Oberski, D. L., & Vermunt, J. K. (2016b). Power and type i error of local fit statistics in multilevel latent class analysis. *Structural Equation Modeling*, *24*(1), 216–229.

Oberski, D. L., van Kollenburg, G. H., & Vermunt, J. K. (2013). A monte carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, *7*(3), 267–279.

Okazaki, S., Campo, S., Andreu, L., & Romero, J. (2014). A latent class analysis of spanish travelers? mobile internet usage in travel planning and execution. *Cornell Hospitality Quarterly*, 1938965514540206.

Oravecz, Z., Faust, K., Batchelder, W. H., & Levitis, D. A. (2015). Studying the existence and attributes of consensus on psychological concepts by a cognitive psychometric model. *The American Journal of Psychology*, *128*(1), 61–75.

Pearlin, L. I., & Johnson, J. S. (1977). Marital status, life-strains and depression. *American sociological review*, *42*, 704–715.

R Core Team. (2012). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/` (ISBN 3-900051-07-0)

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved

from `https://www.R-project.org/`

Reiser, M., & Lin, Y. (1999). A goodness-of-fit test for the latent class model when expected frequencies are small. *Sociological methodology*, *29*(1), 81–111.

Rindskopf, D. (2002). The use of latent class analysis in medical diagnosis. In *Annual meeting of the american statistical association* (pp. 2912–2916).

Rindskopf, D., & Rindskopf, W. (1986). The value of latent class analysis in medical diagnosis. *Statistics in medicine*, *5*(1), 21–27.

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association*, *95*(452), 1143–1156.

Roedelof, A., Bongers, I. L., & van Nieuwenhuizen, C. (2013). Treatment engagement in adolescents with severe psychiatric problems: a latent class analysis. *European child & adolescent psychiatry*, *22*(8), 491–500.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4), 1151–1172.

Rubin, D. B., & Stern, H. S. (1994). Testing in latent class models using a posterior predictive check distribution. In A. Von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research (pp. 420–438)*. Thousand Oaks, CA: Sage Publications Inc.

Sackrowitz, H., & Samuel-Cahn, E. (1999). P values as random variables – expected p values. *The American Statistician*, *53*(4), 326–331.

Salibian-Barrera, M., & Zamar, R. H. (2002). Bootstrapping robust estimates of regression. *Annals of Statistics*, 556–582.

Schaeffer, N. C. (1988). An application of item response theory to the

measurement of depression. *Sociological methodology*, *18*, 271–307.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, *82*(398), 528–540.

van den Bergh, M., Schmittmann, V., & Vermunt, J. (in press). Building latent class trees, with an application to a study of social capital. *Methodology*.

van Kollenburg, G. H., Mulder, J., & Vermunt, J. K. (2015). Assessing model fit in latent class analysis when asymptotics do not hold. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *11*(2), 65–79.

van Kollenburg, G. H., Mulder, J., & Vermunt, J. K. (2017). Posterior calibration of posterior predictive p values. *Psychological Methods*, *22*(2), 382–396.

Vermunt, J. K., & Magidson, J. (2005). Factor analysis with categorical indicators: A comparison between traditional and latent class approaches. In A. K. van der Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences (pp. 41–62).* Mahwah, NJ: Psychology Press.

Vermunt, J. K., & Magidson, J. (2013). Technical guide for Latent GOLD 5.0: Basic, advanced and syntax. *Belmont Massachusetts: Statistical Innovations Inc.*

Vermunt, J. K., & Magidson, J. (2016). Technical guide for Latent GOLD 5.1: Basic, advanced and syntax. *Belmont Massachusetts: Statistical Innovations Inc.*

Von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse

categorical data: Results of a monte carlo study. *Methods of Psychological Research*, *2*(2), 29–48.